DNA-barcoding for the inference of larval community structure of non-biting midges (Chironomidae) from the River Stour, Kent

By

Vikki Louise McConkey

Canterbury Christ Church University

Thesis submitted for the degree of

MSc by Research

2017

Contents

List of Figures and Tables

In order of appearance:

Abstract

Chironomids, or non-biting midges (Chironomidae), belong to the order of true flies (Diptera), and are the most abundant freshwater macroinvertebrates with an estimated 20,000 species worldwide. With the use of the taxonomic identification tool DNA barcoding, it is possible to identify new species and identify specimens without the need for expert knowledge in morphological identification. In this study, DNA barcoding using the mitochondrial (mt) DNA cytochrome oxidase I gene (CO1) was used to identify non-biting midge larvae from the River Stour, Kent. Chironomid larvae were collected by kick sampling at three different sites: Bingley Island (BI), Rheims Way (RW) and Westgate Towers (WT). A total of 93 DNA extractions were carried out including 36 samples from Bingley Island, 30 from Rheims Way and 27 from Westgate Towers. From these, only 33 samples (BI = 8, RW = 12, WT = 13) gave successful PCR amplifications and provided mtDNA sequences for analysis (35.5% success rate). Reasons for the low success rate were discussed; and could be due to the small tissue sample left after the head capsule and final abdominal segments were removed (for morphological analysis), or because of amplification failure due to contamination in the specimen (PCR inhibitors in some samples). Nonetheless, the sequence data obtained was robust enough for analysis. The mtDNA sequences were compared to the public data base GenBank (NCBI) using the Basic Local Alignment Search Tool (BLAST). A similarity score was obtained for each sample, and the first hit was used to provide a 'putative' (tentative or provisional) genus name to the mtDNA sequences from the River Stour. Using phylogenetic methods, the evolutionary relationships of the chironomids obtained in this study were assessed alongside other sequence data obtained from GenBank. Also, genetic diversity and differentiation values among the three sites and GenBank data were obtained. The presence and abundance of chironomids from the River Stour (with their putative genus names) were also used to study the generic richness and diversity in each site, as well as the community structure among the sites. Results showed that it was possible to obtain a putative identification of chironomid larvae, *i.e.* DNA barcoding, as shown by the similarity scores obtained with BLAST. The

phylogenetic analyses also showed the relative similarities of the DNA sequences of chironomids from the River Stour to all other chironomids from GenBank; and also showed the still unexplored non-biting midge phylogenetic relationships. Overall, the three sites in the River Stour were very similar and showed little richness and diversity, although this could be due to the low success obtaining DNA sequences. This study represented the first DNA characterisation of chironomids from the River Stour, and the results could be interpreted as representative of the chironomid larval community. As a first attempt to DNA barcode chironomids from the River Stour, the results and could be useful for monitoring purposes, or as a comparison of biodiversity among other rivers in the region.

Introduction

DNA barcoding as a molecular tool in ecology and evolution

Over a decade ago it was proposed that DNA sequences could be used for quick and reliable species-level identification (Hebert *et al.* 2003, Pečnikar and Buzan 2014). Since then, DNA barcoding, simply defined as the identification of an organism based on molecular data (Hebert *et al.* 2003), has been considered and has become an important tool within the areas of conservation, ecology, evolution and systematics (Schindel and Miller 2005). The two main aims of DNA barcoding have been the discovery of new species and to identify specimens (Schindel and Miller 2005), and it aids and allows the describing, discovering and understanding of biodiversity (Joly *et al.* 2014). DNA barcoding is also especially useful for studying groups of organisms that have been previously neglected in terms of scientific research due to difficulties in the identification and description using morphological features, or because of their high biological diversity (Kress *et al.* 2015). Moreover, the use of DNA barcodes is unfolding the discovery and description of a vast diversity of species formerly unrecognised through the analysis of morphological features alone (Kress *et al.* 2015).

The use of DNA barcoding has continued to increase and has become more popular since 2003 when the first public data base of DNA barcodes was first created (Hebert, Cywinska and Ball 2003). Since 2004 there has been widespread attention and popularity for a global biodiversity inventory through DNA barcoding (Zhang *et al.* 2010). This has been able to successfully present the identification of new species within several groups of taxonomy (Hebert *et al.* 2004; Hajibabaei *et al.* 2006). There are many scientific areas in which DNA barcoding can support research and these are evolutionary and conservation biology, biogeography, biomedicine, epidemiology and ecology (Frézal and Leblois 2008).

Dayrat (2005) stated that ‘Delineating species boundaries correctly is crucial to the discovery of life's diversity because it determines whether or not different individual organisms are members of the same entity.’ This is made possible with the use of DNA barcoding.

In a general sense, DNA barcoding is defined as a molecular tool for the identification of an organism to the level of species or any other taxonomic level using a DNA fragment as a molecular marker (Witt, Threloff and Herbert 2006; Valentini, Pompanon and Taberlet 2009). The technique of DNA barcoding is based on one or several short gene sequences that have been obtained from a specific portion of the genome, and that is used to identify an individual to a preselected taxonomic level.

CO1, the molecular marker of choice

Cytochrome C Oxidase 1 (CO1), also reported as COI or COX1, is a mitochondrial (mt) DNA gene involved in translocating (changing the position of) proteins and electron transport across the membrane (Lunt *et al.* 1996). The mitochondrion is an organelle (membrane bound structure inside cells) present in all eukaryotic cells, including protists, fungi, animals and plants. Mitochondria have their own genome, a circular molecule of DNA containing various genes, like CO1, that code for proteins involved in cellular respiration. Within the CO1 gene, a nucleotide region of 658 base pairs (bp) is frequently used for DNA barcoding (Matz and Nielsen 2005). This DNA fragment represents about 40% of the whole gene and it has a higher rate of genetic variation when compared to other mitochondrial genes (Luo *et al.* 2011). Due to being a mitochondrial gene and its role in metabolism, this gene is relatively well conserved across eukaryotic groups and yet it is still variable enough between species compared to nuclear coded genes (Strüder-Kypke and Lynn 2010). Moreover, this region has been reported to have a precision and reliability close to 100% while identifying individuals in animal studies (Hebert, Cywinska and Ball 2003).

Primarily, the use of DNA barcodes has been the identification of an unknown sample by correctly matching the DNA sequence of a specific genetic marker (*i.e.* the partial CO1 sequence) to a reference sequence from a voucher specimen in a DNA 'library' that has been previously identified and classified taxonomically using morphological characteristics and the same genetic marker (Kress *et al.* 2015). However, DNA barcoding has also been used to

identify new species (Kress *et al.* 2015). This has been possible due to the standardised molecular biology techniques of DNA extraction, Polymerase Chain Reaction (PCR) and DNA sequencing as these techniques increase the speed in which identifying an unknown species is possible (Seifert *et al.* 2007).

In taxonomy, DNA barcoding has been demonstrated to be an effective tool for the successful identification of several vertebrate and invertebrate taxonomic groups (Hajibabaei *et al.* 2007), including Birds (Hebert *et al.* 2004), Fish (Ward *et al.* 2005), Lepidoptera (Hajibabaei *et al.* 2006), Amphibians, Reptiles (Vences *et al.* 2012), Marine organisms (Shander and Willassen 2005), Arachnids (Barrett and Hebert 2005) and Chironomidae (Pfenninger *et al.* 2007).

The DNA barcoding criteria

The development of global standards and coordination of the research in DNA barcoding is driven by an international organisation known as the Consortium for the Barcode of Life (CBOL, http://barcoding.si.edu). The ideal DNA barcoding system should meet several criteria (Valentini, Pompanon and Taberlet 2009), including:

1) The gene region sequenced should be identical or nearly identical among individuals belonging to the same species, but sufficiently different between species. This would allow the identification of a DNA barcode 'gap' for the reliable identification at the species level. A similar scenario should be expected for higher taxonomic levels.
2) The gene region sequenced should be standardised, using the same DNA region for different taxonomic groups.
3) The target DNA region should contain enough phylogenetic information to assign unknown or not yet 'barcoded' species to their taxonomic group.

4) The target DNA region should be robust, with highly conserved priming sites for highly reliable DNA amplification using the Polymerase Chain Reaction (PCR), as well as for DNA sequencing.

5) The target DNA region should be short enough to allow amplification of degraded DNA. Usually, DNA regions longer than 150 bp are difficult to amplify from degraded DNA which is typical of environmental samples, museum and organism that have not been properly preserved for molecular analysis.

CO1 meets all of the DNA Barcoding System criteria listed above and has been accepted as a practical, standardised, species-level DNA barcode for many groups of animals, but not in plants and fungi (Kress *et al.* 2015). In plants and fungi CO1 is not able to be used because mitochondrial genes evolve too slowly, limiting the amount of phylogenetic information and decreasing the accuracy needed for species delimitation (Chase and Fay 2009). For plants, it is recommended that chloroplast genes Ribulose-1, 5-bisphosphate carboxylase oxygenase (rbcL) (Chase *et al.* 1993) and Maturase K (matK) (Jing *et al.* 2011) are recommended for DNA barcoding. For fungi the translation elongation factor 1-α (TEF1α), RNA-polymerase II gene (RPB2), phosphoglycerate kinase (PGK) and DNA topoisomerase I (TOPI) (Stielow *et al.* 2015) are recommended for DNA barcoding.

CO1 thus became the most commonly used genetic marker in eukaryotic organisms (Ekrem, Stur and Hebert. 2010). Moreover, animal identification using DNA barcoding can also be carried out non-invasively from urine, faeces and hair/fur left behind (Valentini, Pompanon and Taberlet 2009), albeit with different levels of success and other limitations, making this a very valuable tool in ecological and evolutionary studies. There are many issues with faecal and urine (endocrinology) DNA sample identification. In faecal DNA barcoding, DNA of host (non-invasive sampling) and biological content (for diet analysis or micro- and macro biome studies) degrades quickly, resulting in low quality DNA that fails to amplify (Piggott 2005, Valentini, Pompanon and Taberlet 2009), and detecting DNA of the host can be difficult due to contamination (Piggott and Taylor 2003). In urine samples, low amounts of cellular DNA of the host is present making DNA barcoding difficult (Hawthorne *et al.* 2009),

and urine can contain DNA from other organisms, also making DNA barcoding difficult due to contamination.

As well as being useful in ecology and evolution, this technique has been applied in the areas of forensic science, biotechnology and the food industry (Valentini, Pompanon and Taberlet 2009, Barcaccia, Lucchin and Cassandro 2015, Mwale *et al.* 2016).

There are several benefits of DNA barcoding as identified by Savolainen *et al.* (2005), and these include:

1) Matching various life stages of the same species.
2) Making the outputs of systematics available to large community of end-users by providing standardised and high-tech identification tools.
3) Providing bio-literacy tool with access for the general public. 4) Easing the burden of identifications from taxonomists by DNA barcoding insects.

Despite there being many uses and advantages of DNA barcoding there are also some issues that are associated with the technique which can provide a less accurate or no species identification. The limitations of DNA Barcoding as stated by Frézal and Leblois (2008) are:

1) A lack of taxa sampling making not enough samples available creating a 'barcoding gap'.

2) Mitochondrial inheritance due to less diversity of mitochondrial DNA.

3) Nuclear copies.

Further to this many of the DNA barcoding reference data bases rely greatly on museum material which is limited with usually low success especially in relation to insects due to degradation of specimens (Meusnier *et al.* 2008).

Specifically, for insects and conservation, insect identification is vital for the management of endangered species, protecting species and controlling invasive species in the ecosystem, and is highly important in ecological research (De Mandal *et al.* 2014). Over 1 million species of insects are present on Earth and represent 50% of all life forms, making insects the most varied and abundant form of life (Pratheepa *et al.* 2014). Using CO1 as the DNA barcoding marker it has been possible to describe numerous species of insects, and the

mitochondrial genome of insects is the most greatly studied molecule, even compared to nuclear genomes (Cameron 2014).

A study by Floyd, Wilson and Hebert (2009) looked into how DNA barcodes help to identify different groups of insects and have examined how it has been possible to increase the knowledge of biodiversity. Insect groups that were included in this work were: Lepidoptera, Diptera, Coleoptera, Hymenoptera, Collembola and Ephemeroptera. As an example, for Lepidoptera they are used as a model group for DNA barcoding studies with this making it possible to link the different life stages of species within the group. Another study by Yusseff-Vanegas and Agnarsson (2017) investigated how DNA barcoding is used within forensic investigations in the Caribbean as groups of insects especially the blow fly (Diptera: Calliphoridae) can help to solve the time and possible cause of death which shows the importance of being able to identify insect species quickly and reliably without using a taxonomic expert.

Chironomids and DNA barcoding

The Order Diptera (the true flies) has around 150,000 described species (Carvalho and Mello-Patiu 2008). The non-biting midges or chironomids (Chironomidae) belong to this order and are one of the most abundant freshwater macroinvertebrates (Ferrington 2008). Globally, non-biting midges are estimated as having 20,000 species (Serra *et al.* 2016) and compose 50% of the total macroinvertebrate community (Coffman and Ferrington 1996) in terms of species richness and abundance (Ferrington 2008). Across Europe it is expected that there is a total of 1262 Chironomidae species (Sæther and Spies 2013). Since 1998, there has been a rapid increase in the number of identified chironomid species (Paasivirta 2014), but despite this increase there has been very little research on Chironomidae taking place in the United Kingdom (UK). In the UK it is estimated that there are around 600 species of non-biting midges (Cranston 2008, Ruse 2013), a substantial increase from 389 known species identified in 1950 (Coe, Freeman and Mattingly 1953), representing about 3% of the global diversity.

Chironomids are important in the maintenance of aquatic ecosystem services, mainly as secondary producers and because they play a fundamental part in energy flow dynamics as an abundant functional feeding group as well as prey items (Nicacio and Juen 2015). Therefore, chironomids, like many other aquatic invertebrates, have been used as indicator species (Ruse and Wilson 1994), mostly as indicators of water quality (Tang *et al.* 2010), because they have the capacity to withstand differing levels of oxygen concentrations, water depth, temperature, pH and salinity (Porinchu and MacDonald 2003), as well as being able to survive in areas with low levels of resources (Failla *et al.* 2015). Despite their importance in research, midges in general (including the non-biting midges – chironomids) are usually described as pests (Failla *et al.* 2016) because they are found in large numbers, they can reproduce rapidly, are easily transported long distances by humans, compete with other organisms for food (Failla *et al.* 2015).

There have been several studies on non-biting midges around the globe, including in places like Antarctica (Kelley *et al.* 2014), Eurasia (Brooks and Langdon 2014), North America (Fortin *et al.* 2015), Central America (Pérez *et al.* 2013), South America (Massaferro *et al.* 2014), Australia (Chang *et al.* 2015) and East Africa (Eggermont and Verschuren 2007). Previous studies using midges have for the majority looked at the abundance and diversity. Ilmonen and Paasivirta (2005) studied the variation in benthic macrocrustacean and insect assemblages in south-west Finland and how the abundance of insect populations varied depending on the habitat type and not the location of the habitat with no significance found statistically. Another study by Theissinger *et al.* (2018) used DNA metabarcoding to observe the abundance and diversity of chironomids in wetlands under different types of mosquito control actions; from this they found 54 chironomid Operational Taxonomic Units (OTUs) across all study sites, and minor effects of the biocide *Bacillus thuringiensis* on the chironomid community composition.

Although the use of chironomids in research and environmental sciences is extensive, it is limited by the ability of the scientists to identify them taxonomically. For chironomids, like in many other insect groups, only adult specimens can be reliably identified to the level of

species because they have more developed and noticeable morphological features than larvae, which can be identified to the level of genus by an expert, and only sometimes to the species level (Ram *et al.* 2014). In contrast, the identification of non-biting midges using DNA barcoding is possible through all life stages (Lima, Floyd and Hanner 2008).

Identification of query sequences in DNA barcoding

There are several methods for assigning species (or other taxonomic level of classification) to DNA sequences. In DNA barcoding, tree-based query identification has been used as a method for assigning a species name (or in this study, a genus name) to the samples, where sequences are considered successfully identified as long as they formed species-specific clusters (Meier *et al.* 2006).

Another DNA barcoding method is through direct sequence comparison where a wide variety of techniques, metrics, criteria and algorithms have been used proposed (Meier *et al.* 2006). One of these methods is the 'best close match' in which the best barcode match of a query is identified, but then only the species name of that barcode is assigned to the query if the barcode is sufficiently similar (if it is not similar, the query remains unidentified) (Meier *et al.* 2006). A threshold of similarity of 95% has been used in intra-specific studies (Meier *et al.* 2006).

A final method for DNA-based identification is based on genetic distance, comparing pairwise intra- and inter-specific genetic distances and detecting a DNA barcode gap between the intra-specific pairwise differences and the inter-specific pairwise differences (Meier *et al.* 2006). Although this last method is very reliable, it depends on a large sample size.

All the methods rely on or use digital repositories of DNA sequences to increase the chances of finding a best match to the query sequence. There are at least two main repositories for DNA sequences used in DNA-barcoding, the Barcode of Life Data System (BOLD) and GenBank, which belongs to the National Center for Biotechnology Information (NCBI). Although BOLD is specifically designed for DNA barcoding, a copy of all sequence

and key specimen data also migrate to NCBI. The NCBI was founded in 1988 as a resource for molecular biology, biochemistry and genetics information with the mission to develop new information technologies to assist in the understanding of fundamental molecular processes that control health and disease (NCBI 2019). GenBank started in 1982 and is a public online database consisting of annotated nucleotide and protein sequences, and supporting bibliographical and biological glossaries for areas of biological significance (Benson *et al.* 2008). Every upload to the GenBank database is recorded and it is identifiable by an assigned accession number, and information includes a nucleotide and/or protein sequence as well as annotations related to the sequences (Benson *et al.* 2008); therefore, it is an important resource in DNA barcoding.

In this study, we focused on the tree-based query identification approach and the best close match approach which do not require large amounts of samples. Also, we combine this with the use of BLAST search results and GenBank data to provide a more reliable identification to determine the community structure of chironomids in the River Stour.

Aims and Objectives

The aims of this study were firstly to identify non-biting midge larvae from the River Stour in Canterbury in Kent, UK through DNA barcoding, and secondly to study the community structure of chironomids in this river. This study is important because overall very little is known about the taxonomic diversity of chironomid larvae in South East (SE) England, and specifically in the River Stour; therefore, a quick and reliable identification tool is needed which does not rely on morphological characters and highly specialised taxonomic experience.

The objectives of this study were:

1) To identify non-biting midges in the River Stour using CO1 DNA barcoding.
2) To compare the genetic composition at various sites along the river and characterise the community structure of non-biting midges based on the DNA barcoding data.

This work represents the first community structure analysis of non-biting midges using DNA barcoding in Kent, and possibly in the South East of England. It also forms the basis for the continued monitoring of chironomid genera or species diversity within the River Stour, and the techniques developed here can easily be expanded for regional studies or further afield.

Materials and Methods

Study site and collection of larvae

The collection of the chironomid larvae took place on the 19th October 2016 and 7th April 2017. The chironomid larvae were obtained from the River Stour in Canterbury, Kent, UK at three different sites (Fig. 1): Bingley Island (BI), Rheims Way (RW) and Westgate Towers (WT). The three sites in the River Stour were selected based on previous studies carried out by the Environment Agency (EA) in 2011, 2012 and 2013 for water quality analysis, and for previous sampling of non-biting midges for morphological identification during an internship at Canterbury Christ Church University (CCCU) in 2013.

The River Stour has the source in the village of Lenham (Environment Agency 2014) and flows into the North Sea at Pegwell Bay (Rogers and Dussart 2004). From Lenham, the river flows to the SE towards Ashford and it is called the Upper Great Stour. Near Hythe to the NW, the river is called the East Stour. The Upper Great Stour and the East Stour confluence near Ashford and the river flows to the NE through Wye, Chilham and Canterbury, where it is called the Great Stour. Near Plucks Gutter, the river flows to the NE towards Reculver and it becomes a tidal river called River Wantsum, but the main River Stour flows from Plucks Gutter to the E towards Sandwich and Pegwell Bay where it is still called River Stour. The length of the River Stour from the source to the mouth is 35 Km (21.7 miles) and has catchment areas in Ashford, Canterbury, Dover, Deal, Ramsgate and Margate (Environment Agency 2013), the second largest in Kent (after the River Medway).

The area surrounding the river and the city of Canterbury is mainly consisting of chalk soils and has been the underlying bed since the Eocene Epoch (Dewey *et al.* 1925, Farrant and Aldiss 2002). The EA monitors the water quality and pollution levels of the River Stour as it is classified as a Nitrate Vulnerable Zone (NVZ). In 2016, the EA water samples and modelling found that the water was polluted but had remained stable from previous monitoring

years, but quality levels did appear to be improving despite the risk of pollution (Environment Agency 2016).

Figure 1. Sample collection sites. Bingley Island (BI), Rheims Way (RW) and Westgate Towers (WT), in Kent County, UK. https://magic.defra.gov.uk/MagicMap.aspx

To obtain chironomid larvae from the river, sediment samples were obtained through kick-sampling and using a pond net (1 mm mesh) following the guidelines of the Freshwater Biological Association (https://www.fba.org.uk/). In each river site, kick-sampling was performed for 3 minutes, kicking the bottom of the river and raising the sediment and rocks

and moving against the water flow along the river. By following this procedure, the sediment and the freshwater invertebrates are caught in the pond net. Large rocks are removed from the net to avoid any damage, but smaller pebbles might be caught in the net. The sediment sample was then placed in a tray for inspection, and the pond net and any small rocks or pebbles were shaken and rinsed with water from the river to collect any invertebrates attached to them. For each site, this procedure was repeated three times moving in parallel lines to cover the width of the river and to obtain sediments from the shallowest (river shores) and deepest (mid-river) parts of the river. The samples were inspected by eye on site, and vertebrates and other easily identifiable and unwanted organisms were removed and returned to the river. The rest of the samples were taken to the laboratory and inspected for the presence and collection of chironomid larvae as described below.

Sorting and Storing Samples

Sediment samples from the three river sites were sought through in the laboratory, and chironomid larvae were initially identified and separated from other insect larvae using a freshwater identification guide (Dobson *et al.* 2012); however, chironomids can be easily distinguished from other freshwater invertebrates and identified by eye because of their overall shape, size, colour and movement. Chironomid larvae were stored in screw-cap tubes, clearly labelled according to their site, containing 95% Ethanol and kept at -20 °C until DNA extraction.

DNA Extraction

Before DNA extraction and to aid future morphological identification of larvae (Fig. 2), all individual larvae were labelled, and the head capsule and posterior abdominal segments were removed from all specimens and stored in 95% Ethanol at -20 °C. DNA from the remaining abdominal segments was extracted using the GeneJet DNA purification kit following the instructions manual (ThermoFisher Scientific).

The labelling for individual larvae included: 1) name of river – Stour (ST); 2) name of sampling site – Bingley Island (BI), Rheims Way (RW) and Westgate Tower (WT); 3) sample number. For example, the first larva from Bingley Island was labelled STBI1 (for Stour, Bingley Island, and sample #1).

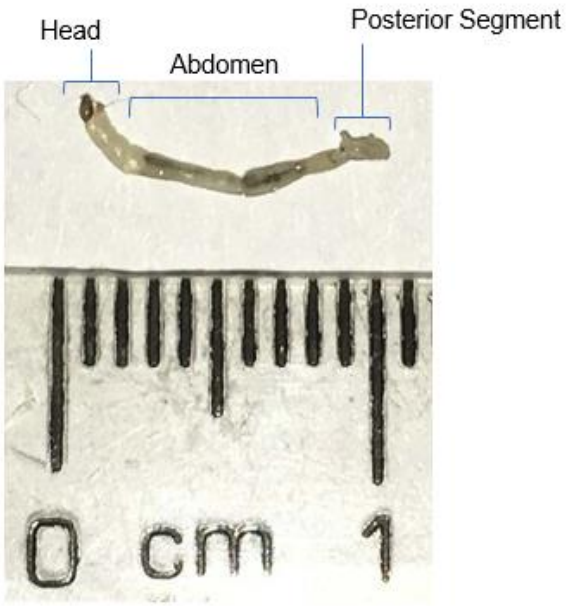


Figure 2. Photo of a chironomid larva taken before removal of head and posterior segments.

Polymerase Chain Reaction (PCR)

PCR is a molecular biology method used to make copies of ('amplify') specific DNA regions/segments of interest by a process that shares similarities with DNA replication (Mullis *et al.* 1987). In PCR, the DNA of the organism of interest is mixed with several reagents (e.g. Primers and DNA Polymerase) in a plastic 'PCR' tube or plate; the amplification of the target DNA region takes place following six steps (Initialisation, Denaturation, Annealing, Elongation, Final Elongation and Final Hold) that involve the change of temperature using a thermo-cycler (a piece of equipment that can change temperature of the tubes very quickly and hold it for a

pre-determined amount of time) (Mullis *et al.* 1987). In PCR, the primers are designed short single strands of DNA (c. 20 bl long) that bind to specific target regions in the target DNA and allow the DNA polymerase to make copies of that region (Jaric *et al*. 2013). Primers used can be either 'forward' or 'reverse'; a forward primer binds to the 5' end of the leading (template) DNA strand and begins to elongate until reaching the 3' end, while the reverse primer binds to the 3' end of the lagging DNA strand and begins to elongate until reaching the 5' end (Schoenbrunner *et al*. 2017). Another important component for the PCR process is the DNA polymerase, or 'Taq Polymerase', 'Taq Pol' or simply 'Taq', which was originally isolated from the bacterium *Thermus aquaticus* in 1976 (Chien, Edgar and Trela 1976, Saiki *et al.*1985). The *Thermus aquaticus* bacterium was discovered in hydrothermal vents and is an extreme thermophile meaning that the DNA polymerase is able to withstand the high temperatures needed as part of PCR (Saiki *et al.* 1988).

The primers used in PCR were the forward primer LCO1490 (5'-ggtcaacaaatcataaagatattgg-3') and the reverse primer HCO2198 (5'-taaacttcagggtgaccaaaaaatca-3') (Folmer *et al.* 1994). In the code names for the primers, L and H refer to light and heavy DNA strands, CO refers to cytochrome oxidase subunit I, and the numbers (1490 and 2198) refer to the position of the *D. yakuba* 5' nucleotide (Folmer *et al.* 1994). These 'universal' primers were created at the time after the discovery of around 230 invertebrate species from deep-sea hydrothermal vents and cold-water sulphide or methane seep communities (Tunnicliffe 1991). For the candidate 'universal' primers to be designed, published DNA sequences for the following species were used (Folmer *et al.* 1994): blue mussel, *Mytilus edulis*; fruit fly, *Drosophila yakuba*; honeybee, *Apis mellifera*; mosquito, *Anopheles gambiae*; brine shrimp, *Artemia franciscana*; nematodes, *Ascaris suum* and *Caenorhabditis elegans*; sea urchin, *Strongylocentrotus purpuratus*; carp, *Cyprinus carpio*; frog, *Xenopus laevis*; chicken, *Gallus gallus*; mouse, *Mus musculus*; cow, *Bos taurus*; fin whale, *Balaenoptera physalus*; and human, *Homo sapiens*. The LCO1490 and HCO2198 primers were chosen because their amplification was consistent to 710-bp from the broadest

array of invertebrates. Since then, these primers have been used successfully in many other studies on insects, including chironomids (Folmer *et al.* 1994, Clary and Wolstenholme 1985).

PCRs were done in 50 µL of final volume, mixing 25 µL of DreamTaq Green PCR Master Mix (2X) containing DreamTaq DNA Polymerase, 2X DreamTaq Green buffer, dNTPs, and 4 µM $MgCl_2$ (ThermoFisher Scientific), 18 µL of molecular grade water, 2.5 µL of Forward Primer LCO1490, 2.5 µL of Reverse Primer HCO2198, and 2 µL of genomic DNA sample. All PCRs were done in sterile 200 µL microcentrifuge tubes, PCR reagents and DNA were added using sterile microtips, the PCR tubes were vortexed for 2 seconds to ensure that all contents were thoroughly mixed, and the PCR tubes were briefly centrifuged for 5 seconds to bring all contents to the bottom of the tube. Negative controls (adding 2 µL of water instead of DNA) were done in each PCR to ensure that no contamination was present during amplification.

The following protocol of Folmer *et al.* (1994) was used for the amplification of 710 base pairs (bp) of the mtDNA gene CO1: an initial denaturation step of 5 minutes at 95 °C, followed by 35 cycles of 1 minute at 95 °C for denaturation, 1 minute at 40 °C for annealing, and 1.5 minutes at 72 °C for elongation, with a final elongation step for 7 minutes at 72 °C. PCRs were stored at 4 °C until downstream analysis.

Gel Electrophoresis

Gel Electrophoresis is the process where DNA molecules are separated through an agarose gel according to their size when an electric current is applied. The smaller the size of the DNA the quicker the strands move through the gel. As the DNA separates by size, it moves through the gel forming clusters (bands) of similar size which can be visualized under UV light using a transilluminator when the DNA is stained with an intercalant molecular (*e.g.* Ethidium bromide or SYBRsafe®, Invitrogen). In this case, the PCR products would have an expected size of 710 bp (Folmer *et al.* 1994).

Electrophoresis was performed in 1% Agarose gels (0.4 g Agarose in 40 ml 1X TAE buffer pH 8.0). For each sample, 5 µL of PCR were added to the wells alongside the

GeneRuler 1 kb DNA Ladder (ThermoFisher Scientific). No loading dye was used because it was already premixed in the PCR Master Mix. Agarose gels were run at 80 V for 45 minutes in 1X TAE buffer pH 8.0 and then photographed under UV light using a Gel Doc XR+ Gel Documentation System (Bio-Rad). PCR products were detected by comparing their sizes with the banding pattern produced by DNA ladder. No negative controls amplified; therefore, it was assumed that all PCRs were the results of the amplification of chironomid DNA and not contamination. If no amplification was detected in a sample, the PCR was repeated; however, most samples amplified in the first trial.

PCR Purification and DNA sequencing

PCR products were purified using the GeneJet PCR purification kit following the instructions manual (ThermoFisher Scientific). Purified PCR products were visualised in 1% Agarose gels as described above in Gel Electrophoresis. For DNA sequencing of CO1, purified DNA samples were packaged and sent to DBS Genomics at Durham University following instructions from the sequencing facility. Samples were sequenced with the Sanger sequencing method using the forward primer. DNA sequences were checked visually for quality. Samples that failed to produce a reliable sequence were re-amplified and sent for sequencing.

Sanger sequencing is a chain termination method for determining the nucleotide sequence of DNA developed by Sanger *et al.* (1977). It consists of three steps, including 1) the generation of DNA fragments of varying lengths using a DNA polymerase, each terminated with a labelled nucleotide, 2) the separation of the DNA fragments using capillary gel electrophoresis, and 3) the detection of the labelled nucleotides using a laser which generates a fluorescent peak of different wavelength for each labelled nucleotide, recorded by a computer as a chromatogram.

DNA Sequence Data Analysis

All chironomid larvae DNA sequences were run through the GenBank data base. To search for matching sequences in GenBank, the Basic Local Alignment Search Tool (BLAST) was used. BLAST is a suit of bioinformatic programs provided by the NCBI. BLAST was designed by Altschul *et al.* (1990) and it is one of the most used bioinformatic programs for searching DNA sequences and doing rapid sequence comparisons (Altschul *et al.* 1990, Casey 2005). This tool finds regions of similarity between biological sequences (nucleotide or protein) by comparing an alignment of a 'query' sequence against 'subject' sequences in the GenBank data base and calculates the statistical significance of the matches (Wheeler and Bhagwat 2007). The similarity search generates several outputs, including:

- A list of all hits with a link to the actual GenBank file.
- Maximum Score (Max Score): the highest alignment score of a set of aligned segments from the same subject (data base) sequence. The score is calculated from the sum of the match rewards and the penalties given during the alignment of the sequences due to mismatch, gap open and extend independently for each segment. This normally gives the same sorting order as the E Value.
- Total Score: the sum of alignment scores of all segments from the same subject sequence.
- Query Cover: the percentage of a given input sequence (the query) that has aligned with a sequence on the GenBank data base (the subject). A 100% query cover means that the query sequence has aligned completely to the subject. This is calculated over all segments as with the Total Score.
- Expect value (E value): a parameter used to represent the number of matches (hits) that can be expected to be seen by chance when searching the NCBI BLAST data base. The closer an E value is to 0 the more significant the matches are.

- Percentage Identity (Per. Identity): the measurement of likeness between two DNA sequences. The higher the percentage score the more closely related the sequences are expected to be.
- Accession: the accession number is a unique identifier given to a biological polymer sequence (DNA, RNA or protein) when it is submitted to GenBank.

BLAST was used to identify the 'putative' genus name based on the best matching subject, the 'first hit' or best match. The second hit on BLAST was also recorded; and was simply used for comparisons purposes and to give validity to the first hit – if both hits in BLAST resulted in the same genus name, then the putative name for the DNA sequence would be very reliable (of course also considering the percent similarity). However, if the second hit returned a different genus name, then the similarity scores could be checked; if the similarity of the second hit was much lower, then confidence could be placed on the first hit, but if the similarity scores of first and second hits were very similar then that putative name would not be very reliable. BLAST results for each mtDNA sequence obtained from the Stour were recorded, and the Max Score, Total Score, Query Cover, E value and Per. Identity were averaged by site (Bingley Island, Rheims Way and Westgate Towers) and by genus.

The DNA sequences of the 1st and 2nd hits from GenBank were added to a BioEdit version 7.2.6 (Hall 1999) file alongside an outgroup (*Aedes albopictus*; Diptera, Culidae; Wilkerson *et al.* 2015). Other DNA sequences belonging to Chironomidae were also included in the BioEdit file. The full set of DNA sequences (*i.e.* this study, 1st and 2nd hits in BLAST, other Chironomids from GenBank, and the outgroup) were aligned using the ClustalW Alignment Tool in BioEdit software. Sequence alignment is the arrangement of DNA sequences by regions (nucleotide sites in the DNA sequence) that are similar due to evolutionary or structural relationships between the sequences (Mount 2004), and it is a necessary step while building phylogenies or comparing sequences against data bases.

A Neighbour-Joining phylogenetic tree using all the sequences from the Stour and GenBank, with 500 bootstraps and rooted to the outgroup, was created in MEGA version 7 (Kumar, Stecher and Tamura 2015). The Neighbour-Joining method was created in 1987 by

Naruya Saitou and Masatoshi Nei, and it is a clustering method used in making phylogenetic trees through an algorithm by finding pairs of operational taxonomic units (OTUs or neighbours). This method relies on the estimation of a distance matrix among all pairs of sequences and starts with an unresolved tree (*i.e.* a star shaped tree). The algorithm then searches for the pair of taxa with the shortest genetic distance, which are joined to create a new node connected to the centre of the rest of the unresolved tree. The algorithm then generates a new distance matrix from all sequences to this new node. The algorithm is repeated, using the new node and the distances calculated in the previous step. All possible topologies are examined bioinformatically, and the topology that shows the least evolutionary change is given as the final tree (Saitou and Nei 1987). The phylogenetic tree was used as a method for estimating the relationship of the mtDNA sequences obtained in this study with those already present in GenBank.

The software DnaSP version 5 (Rozas 2009) was used for the DNA sequence analysis of the sample divided into the three sites and GenBank, or into the genera identified with BLAST. Data sets were defined to compare sequence polymorphism between Stour and GenBank samples and to compare populations. DnaSP was used to estimate the number of parsimony informative sites, the number of segregating sites, the number of haplotypes, the haplotype diversity and the nucleotide diversity (explained below).

Parsimony informative sites are those nucleotide sites in the DNA sequences that contain at least two types of nucleotides (polymorphic), and at least two of them occur with a minimum frequency of two. Segregating sites are simply those sites that are polymorphic, irrespective of the frequency of the polymorphism in the sample. The number of haplotypes is simply the variants of a gene or gene fragment (the CO1 fragment in this study) which may differ in just one nucleotide base. Haplotype diversity is the probability that two alleles sampled randomly are different. The nucleotide diversity is the average number of nucleotide differences in each site in pairwise comparisons found among DNA sequences. The haplotype and nucleotide diversity values are used to describe the genetic diversity within populations (De Jong *et al.* 2011).

Community composition in the River Stour

To study the community composition of chironomids in the River Stour, a phylogenetic approach and a traditional diversity index were used. To look at the diversity of haplotypes across the sites, haplotype phylogenetic networks per genus were used.

Haplotype phylogenetic networks for the genera identified in the River Stour (including river and GenBank samples) were generated using Network version 5.0.0.3 (Fluxus Technology Ltd. 2015). A phylogenetic network is a graph or reticulate pattern constructed using the polymorphic (segregating) sites in the DNA sequences, where taxa are represented by nodes and their evolutionary relationships are represented by edges, and it is used to visualise phylogenetic relationships (Huson and Bryant 2006). In phylogenetic networks, the size of each node represent the number of haplotypes present (the larger the node the more haplotypes are present), and the internodes represent the relationships among the haplotypes, and the length of the internodes can be used to represent the number of mutations between haplotypes; however in this study, due to the very long distances among haplotypes (high number of mutational steps between the samples and the GenBank data), the internodes were modified to only show the phylogenetic relationships, not the number of mutations.

Using the software PAST version 3.25 (Hammer, Harper and Ryan 2001) and the number of samples per genus per site, the genus richness and the genus diversity index were calculated. In this study, genus richness is simply the number of putative genera identified in total and per site. The genus diversity index for the River Stour and per site was obtained using Simpson's diversity index (D) as implemented in PAST:

$$D = 1 - \Sigma \left(\frac{n}{N}\right)^2$$

Where n = the total number of organisms of a species (in this case, the total number of sequences identified to a genus), and N = the total number of organisms of all species (in this case, the total number of mtDNA sequences).

In Simpson’s diversity index, a score of 0 determines that there is no diversity for a site whereas a score of 1 shows that there is an infinite species/genus diversity (Simpson 1949). The Simpson's diversity index is used in ecological surveys to calculate the diversity of species based on species richness and their abundance in a given area (Simpson 1949). To calculate Simpson’s diversity index, the putative genus and the abundance per site were used. Although the data obtained was not strictly quantitative, kick-sampling was done in a methodical way; therefore, the abundance data was considered semi-quantitative.

The abundance data per site and generic identification were used to obtain the Bray-Curtis similarity matrix, and the distance values among sites were used to construct an Unweighted Pair Group Method with Arithmetic Mean (UPGMA) dendrogram (cluster analysis). Bray-Curtis Similarity Index is a statistical equation that is used to assess the similarity between sites depending the species found; this is displayed as a 0 or 1, where a distance = 0 would mean that the sites contain all the same species, and a distance = 1 would mean that there are no species shared between sites (Bray and Curtis 1957). UPGMA is a distance-matrix method used to construct phylogenetic trees by calculating the pairwise genetic distances between taxa (Hua *et al.* 2017). Cluster analysis is grouping, or clustering species together based on how similar they are and is used in building phylogenetic networks and trees (Bailey 1994).

Results

DNA sequences and BLAST

In total, 93 DNA extractions were carried out, including 36 samples from Bingley Island, 30 samples from Rheims Way and 27 from Westgate Towers. From these, only 33 samples (BI = 8, RW = 12, WT = 13) gave successful PCR amplifications and provided good quality DNA sequences for analysis (35.5% success). All other chironomid samples were retried for CO1 amplification and sequencing but failed. All DNA sequences will be uploaded to GenBank upon publication of results, and mtDNA sequences as well as samples are stored at CCCU. The 33 DNA sequences from the River Stour were compared against GenBank using BLAST, and the first and second hits were recorded. A total of 542 DNA sequences were aligned and analysed (33 from the River Stour and 509 from GenBank). All DNA sequences obtained here had a minimum length of 600 base pairs (bp).

As explained above, the allocation of putative genera for chironomid samples from the River Stour was based on the first hit (*i.e.* the best match) against the GenBank data base using BLAST. All DNA sequences from the River Stour were identified with BLAST as belonging to nine different genera in the family Chironomidae (see Appendix Table S1), and average GenBank identity from BLAST results by site and by genus are shown in Table 1. Most of the samples from the River Stour (30 mtDNA sequences) had hits in GenBank with genus and species names recorded; however, the other three mtDNA sequences had hits in GenBank with only genus name recorded; therefore, all analyses in this study were done at the genus level rather than at the species level. In 16 mtDNA sequences, the first and second hits returned the same genus names in GenBank. The other 17 mtDNA sequences, all identified in the first hit as *Microtendipes*, the second hit in GenBank only had the family level classification (Chironomidae) and no details below this taxonomic level.

Table 1. Average GenBank identity (using BLAST) of the chironomid larvae from the River Stour per site and per genus.

	Max Score	Total Score	Query Cover (%)	E value	Per. Identity
By site					
Bingley Island (BI)	929.71	929.71	93.60	< 0.001	93.30
Rheims Way (RW)	886.62	886.62	93.50	< 0.001	92.50
Westgate Towers (WT)	997.87	997.87	92.50	< 0.001	95.60
By genus					
Microtendipes	925.00	925.00	93.80	< 0.001	93.40
Paratendipes	770.75	770.75	92.80	< 0.001	89.00
Cricotopus	1068.67	1068.67	94.00	< 0.001	97.67
Chironomus	1040.00	1040.00	95.00	< 0.001	97 12
Phaenopsectra	1122.00	1122.00	92.00	< 0.001	99 13
Tvetenia	1034.67	1034.67	88.30	< 0.001	98.67
Orthocladius	889.50	889.50	93.00	< 0.001	92.00
Rheocricotopus	984.00	984.00	92.00	< 0.001	94.00
Paratanytarsus	935.00	935.00	92.00	< 0.001	93.00

Overall, the highest identity obtained was 99% (including two matches to *Tvetenia*, two to *Cricotopus* and one *Phaenopsectra*), with a mean identity percentage of 94% for all sample sequences (see Appendix Table S1). The lowest identity obtained was 89% (four matches to *Paratendipes*). However, in all cases the E value was below 0.001 (Table 1); therefore, there is high confidence that the taxonomic identification of the mtDNA sequences to the GenBank

data base is correct. The river site with the highest percent identity to sequences in GenBank was Westgate Towers, while the lowest was Rheims Way. Among the identified putative genera, *Phaenopsectra* had the highest average percent identity, while *Paratendipes* had the lowest average percent identity (Table 1).

DNA sequence polymorphism

In total, 542 DNA sequences were analysed, and the total data base consisted of 33 DNA sequences from the River Stour and 509 DNA sequences from GenBank. The number of haplotypes and haplotype diversity for all sites including Stour and GenBank are shown in Table 2.

For the whole data combined (Stour and GenBank together), a total of 335 haplotypes were identified. Out of the 33 DNA sequences from the River Stour, there were 25 haplotypes. The number of haplotypes was very similar in WT and RW, while BI showed the lowest number of haplotypes. As expected, the GenBank sample showed higher haplotype and nucleotide diversity than the full River Stour sample or than any specific site within the River Stour. Important differences were found in nucleotide diversity values when comparing the GenBank data set with all the River Stour sites. However, considering the difference in sample size, there was a comparable haplotype diversity between the GenBank data set, the full River Stour, WT and RW. In BI, only one genus (*Microtendipes*) and five haplotypes were found across eight mtDNA sequences, while RW had 12 sequences with 11 haplotypes and five genera, and WT had 13 sequences with haplotypes and eight genera. BI also had the lowest average number of nucleotide differences than any other group of sequences.

Table 2. DNA Polymorphism of chironomids from the River Stour and GenBank.

	No. of Sequences	No. of Sites	Number of Haplotypes (h)	Haplotype (gene) diversity (Hd)	Nucleotide Diversity (π)	Average number of nucleotide differences (K)
All data (Stour and GenBank)	542	638	335	0.994	0.189	91.799
All Stour	33	638	25	0.956	0.135	74.723
All GenBank	509	638	310	0.994	0.189	92.028
Bingley Island (BI)	8	638	5	0.786	0.006	3.929
Westgate Towers (WT)	13	638	12	0.987	0.165	91.462
Rheims Way (RW)	12	638	11	0.985	0.131	72.727

Phylogenetic analysis

All chironomid sequences were used to build the Neighbour-Joining (NJ) tree (Fig. 3). *Aedes albopictus* was the outgroup and served as a reference to the other nodes (descendants) and root the tree.

As expected, 'putative' genera of chironomids from the River Stour grouped together with the corresponding sequences collected from GenBank (Fig. 3). Although it was not the

intention of this study to infer the phylogeny of chironomids, the NJ tree showed the phylogenetic relationships among all samples (Stour + GenBank) and roughly recovered the subfamily level of classification expected from morphological classifications. For example, the genera *Cricotopus*, *Orthocladius*, *Rheocricotopus* and *Tvetenia* (all belonging to the subfamily Chironominae) clustered together with corresponding genera from GenBank. However, there was little bootstrap support at the base of the tree and in the most basal nodes, and higher bootstrap support on the higher nodes.

Overall, chironomids from the River Stour were DNA barcoded reliably to the genus level, but the phylogenetic relationships among genera remains unresolved. Also, a few genera appeared to be paraphyletic (*i.e.* descending from a common evolutionary ancestor, but the group does not include all the descendants), including: *Rheocricotopus*, *Thienemanniella*, *Corynoneura*, *Limnophyes*, *Metriocnemus*, *Micropsectra*, *Paratanytarsus*, and *Tanytarsus*.

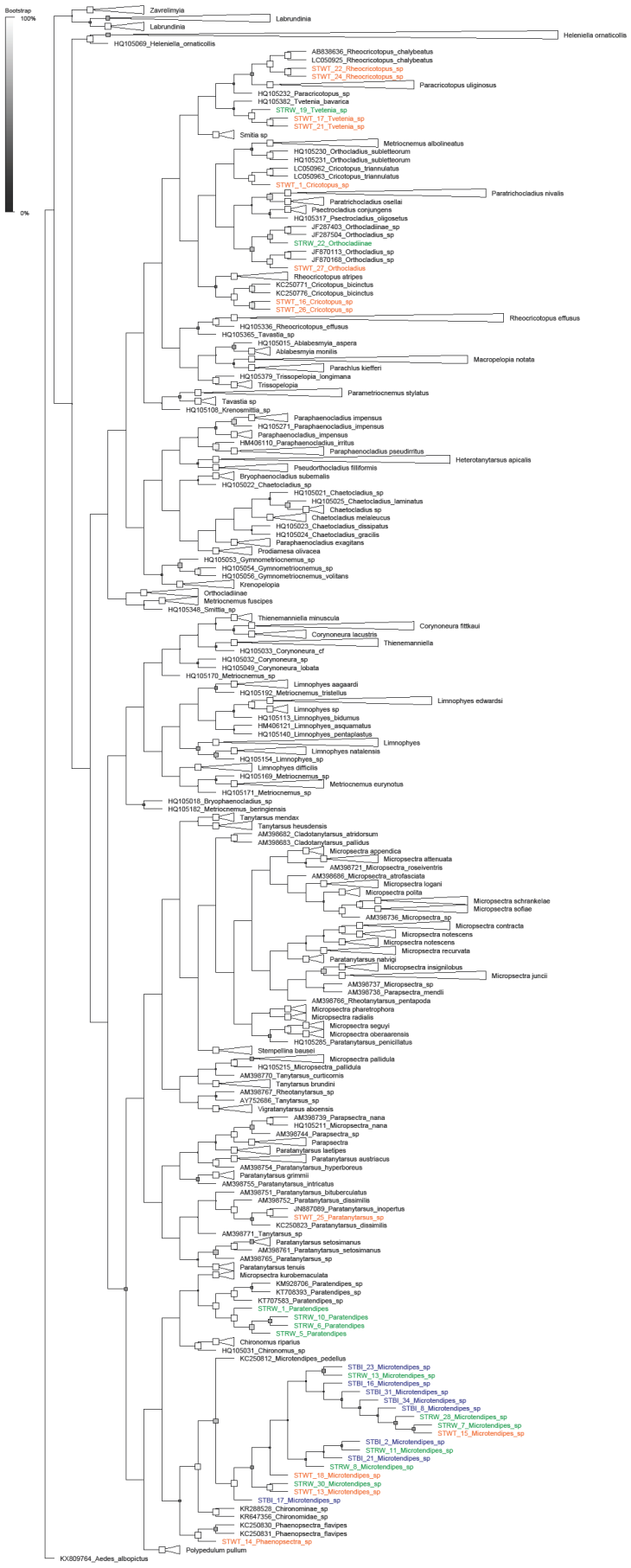
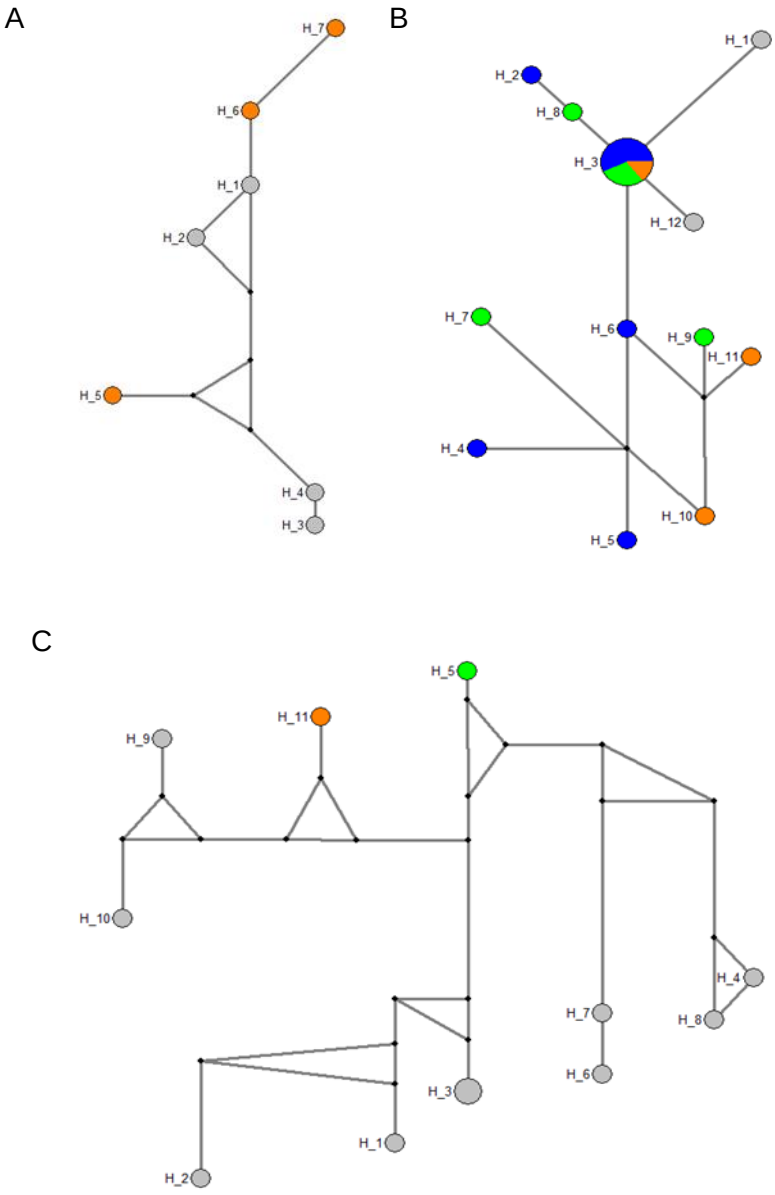

Bootstrap
100%
0%
Zavrelimyia
Labrundinia
Labrundinia
Heleniella ornaticollis
HQ105069_Heleniella_ornaticollis
AB838636_Rheocricotopus_chalybeatus
LC050925_Rheocricotopus_chalybeatus
STWT_22_Rheocricotopus_sp
STWT_24_Rheocricotopus_sp
Paracricotopus uliginosus
HQ105232_Paracricotopus_sp
HQ105382_Tvetenia_bavarica
STRW_19_Tvetenia_sp
STWT_17_Tvetenia_sp
STWT_21_Tvetenia_sp
Smittia sp
Metriocnemus albolineatus
HQ105230_Orthocladius_subletteorum
HQ105231_Orthocladius_subletteorum
LC050962_Cricotopus_triannulatus
LC050963_Cricotopus_triannulatus
STWT_1_Cricotopus_sp
Paratrichocladius nivalis
Paratrichocladius osellai
Psectrocladius conjungens
HQ105317_Psectrocladius_oligosetus
JF287403_Orthocladiinae_sp
JF287504_Orthocladius_sp
STRW_22_Orthocladiinae
JF870113_Orthocladius_sp
JF870168_Orthocladius_sp
STWT_27_Orthocladius
Rheocricotopus atripes
KC250771_Cricotopus_bicinctus
KC250776_Cricotopus_bicinctus
STWT_16_Cricotopus_sp
STWT_26_Cricotopus_sp
Rheocricotopus effusus
HQ105336_Rheocricotopus_effusus
HQ105365_Tavastia_sp
HQ105015_Ablabesmyia_aspera
Ablabesmyia monilis
Macropelopia notata
Parachlus kiefferi
HQ105379_Trissopelopia_longimana
Trissopelopia
Parametriocnemus stylatus
Tavastia sp
HQ105109_Krenosmittia_sp
Paraphaenocladius impensus
HQ105271_Paraphaenocladius_impensus
Paraphaenocladius_impensus
HM406110_Paraphaenocladius_irritus
Paraphaenocladius pseudirritus
Heterotanytarsus apicalis
Pseudorthocladius filiformis
Bryophaenocladius subernalis
HQ105022_Chaetocladius_sp
HQ105021_Chaetocladius_sp
HQ105025_Chaetocladius_laminatus
Chaetocladius sp
Chaetocladius melaleucus
HQ105023_Chaetocladius_dissipatus
HQ105024_Chaetocladius_gracilis
Paraphaenocladius exagitans
Prodiamesa olivacea
HQ105053_Gymnometriocnemus_sp
HQ105054_Gymnometriocnemus_sp
HQ105056_Gymnometriocnemus_volitans
Krenopelopia
Orthocladiinae
Metriocnemus fuscipes
HQ105348_Smittia_sp
Thienemanniella minuscula
Corynoneura fittkaui
Corynoneura lacustris
Thienemanniella
HQ105033_Corynoneura_cf
HQ105032_Corynoneura_sp
HQ105049_Corynoneura_lobata
HQ105170_Metriocnemus_sp
Limnophyes aagaardi
HQ105192_Metriocnemus_tristellus
Limnophyes edwardsi
Limnophyes sp
HQ105113_Limnophyes_bidumus
HM406121_Limnophyes_asquamatus
HQ105140_Limnophyes_pentaplastus
Limnophyes
Limnophyes natalensis
HQ105154_Limnophyes_sp
Limnophyes difficilis
HQ105169_Metriocnemus_sp
Metriocnemus eurynotus
HQ105171_Metriocnemus_sp
HQ105018_Bryophaenocladius_sp
HQ105182_Metriocnemus_beringiensis
Tanytarsus mendax
Tanytarsus heusdensis
AM398682_Cladotanytarsus_atridorsum
AM398683_Cladotanytarsus_pallidus
Micropsectra appendica
Micropsectra attenuata
AM398721_Micropsectra_roseiventris
AM398686_Micropsectra_atrofasciata
Micropsectra logani
Micropsectra polita
Micropsectra schrankelae
Micropsectra sofiae
AM398736_Micropsectra_sp
Micropsectra contracta
Micropsectra notescens
Micropsectra notescens
Micropsectra recurvata
Paratanytarsus natvigi
Micropsectra insignilobus
Micropsectra juncii
AM398737_Micropsectra_sp
AM398735_Parapsectra_mendli
AM398766_Rheotanytarsus_pentapoda
Micropsectra pharetrophora
Micropsectra radialis
Micropsectra seguyi
Micropsectra oberaarensis
HQ105265_Paratanytarsus_penicillatus
Stempellina bausei
Micropsectra pallidula
HQ105215_Micropsectra_pallidula
AM398770_Tanytarsus_curticornis
Tanytarsus brundini
AM398767_Rheotanytarsus_sp
AY752686_Tanytarsus_sp
Vigratanytarsus aboensis
AM398739_Parapsectra_nana
HQ105211_Micropsectra_nana
AM398744_Parapsectra_sp
Parapsectra
Paratanytarsus laetipes
Paratanytarsus austriacus
AM398754_Paratanytarsus_hyperboreus
Paratanytarsus grimmii
AM398755_Paratanytarsus_intricatus
AM398751_Paratanytarsus_bituberculatus
AM398752_Paratanytarsus_dissimilis
JN887089_Paratanytarsus_inopertus
STWT_25_Paratanytarsus_sp
KC250823_Paratanytarsus_dissimilis
AM398771_Tanytarsus_sp
Paratanytarsus setosimanus
AM398761_Paratanytarsus_setosimanus
AM398765_Paratanytarsus_sp
Paratanytarsus tenuis
Micropsectra kurobemaculata
KM526706_Paratendipes_sp
KT708393_Paratendipes_sp
KT707583_Paratendipes_sp
STRW_1_Paratendipes
STRW_10_Paratendipes
STRW_6_Paratendipes
STRW_5_Paratendipes
Chironomus riparius
HQ105031_Chironomus_sp
KC250812_Microtendipes_pedellus
STBI_23_Microtendipes_sp
STRW_13_Microtendipes_sp
STBI_16_Microtendipes_sp
STBI_31_Microtendipes_sp
STBI_34_Microtendipes_sp
STBI_8_Microtendipes_sp
STRW_28_Microtendipes_sp
STRW_7_Microtendipes_sp
STWT_15_Microtendipes_sp
STBI_2_Microtendipes_sp
STRW_11_Microtendipes_sp
STBI_21_Microtendipes_sp
STRW_8_Microtendipes_sp
STWT_18_Microtendipes_sp
STRW_30_Microtendipes_sp
STWT_13_Microtendipes_sp
STBI_17_Microtendipes_sp
KR288528_Chironominae_sp
KR647356_Chironomidae_sp
KC250830_Phaenopsectra_flavipes
KC250831_Phaenopsectra_flavipes
STWT_14_Phaenopsectra_sp
Polypedulum pullum
KX809764_Aedes_albopictus
3.0

Figure 3 (continued from previous page). Neighbour-Joining phylogenetic tree of chironomids from the River Stour and GenBank, rooted with the outgroup *Aedes albopictus*. Squares on nodes represent bootstrap support values (500 bootstraps), where large squares and white fill represent high bootstrap support (100 %) and small squares and black fill represent low bootstrap support (0 %). Relevant zoomed in sections of the tree can be observed in detail in Appendix Fig. S1.

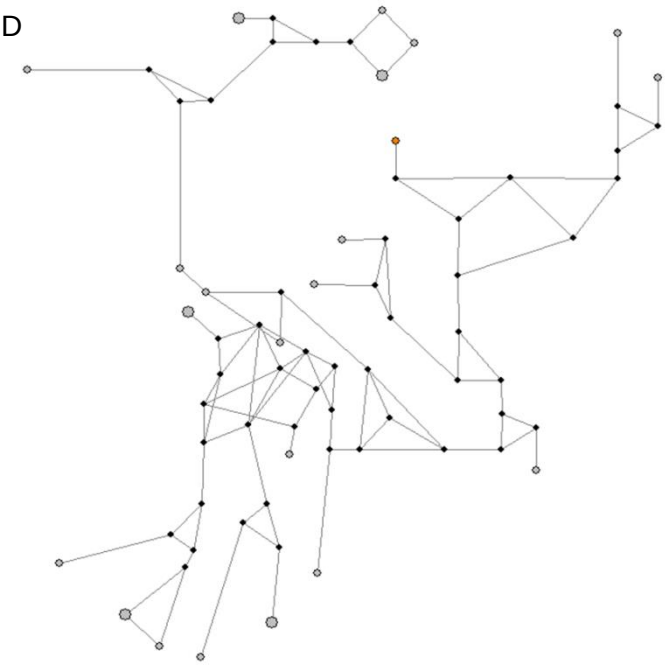
Community composition in the River Stour

The phylogenetic networks (Fig. 4) showed the evolutionary relationships for each of the different genera found in the River Stour and GenBank data. In all cases, the haplotypes from GenBank were different from the ones from the River Stour (*i.e.* the haplotypes were not shared). Overall, this shows that the River Stour had chironomids that were genetically different from, but closely related to, the chironomids reported in GenBank.

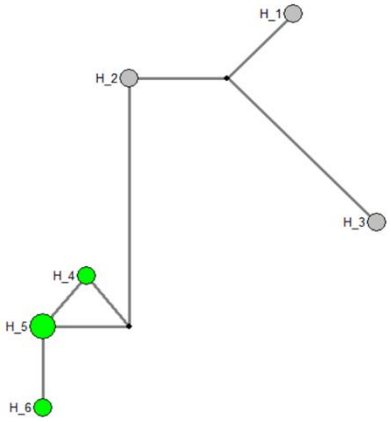
For *Cricotopus* (Fig. 4A) there were three haplotypes in WT, with two of the haplotypes being more similar than the other. *Microtendipes* (Fig. 4B) was the most diverse genus with ten haplotypes found in all three river sites; one haplotype was present in all three river sites; BI had five haplotypes; RW had four haplotypes and WT had three haplotypes. *Orthocladius* (Fig. 4C) had two haplotypes in the River Stour, one in RW and one in WT. *Paratanytarsus* (Fig. 4D) had one haplotype present in WT; however, there were many GenBank haplotypes and all distantly related to the one from the River Stour. *Paratendipes* (Fig. 4E) had three haplotypes all only present in RW. *Phaenopsectra* (Fig. 4F) had only one haplotype present in WT. *Rheocricotopus* (Fig. 4G) had two haplotypes present for WT. *Tvetenia* (Fig. 4H) had two haplotypes present, one in RW and the other one in WT.


A
H_7
H_6
H_1
H_2
H_5
H_4
H_3
B
H_1
H_2
H_8
H_3
H_12
H_7
H_6
H_9
H_11
H_4
H_10
H_5
C
H_5
H_11
H_9
H_10
H_4
H_7
H_8
H_6
H_3
H_1
H_2

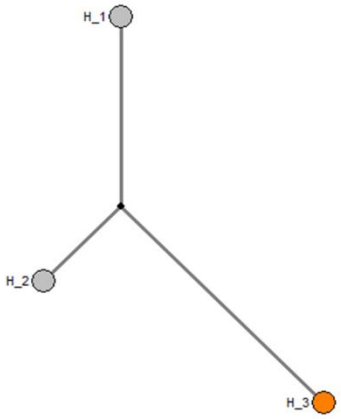
D



E



F



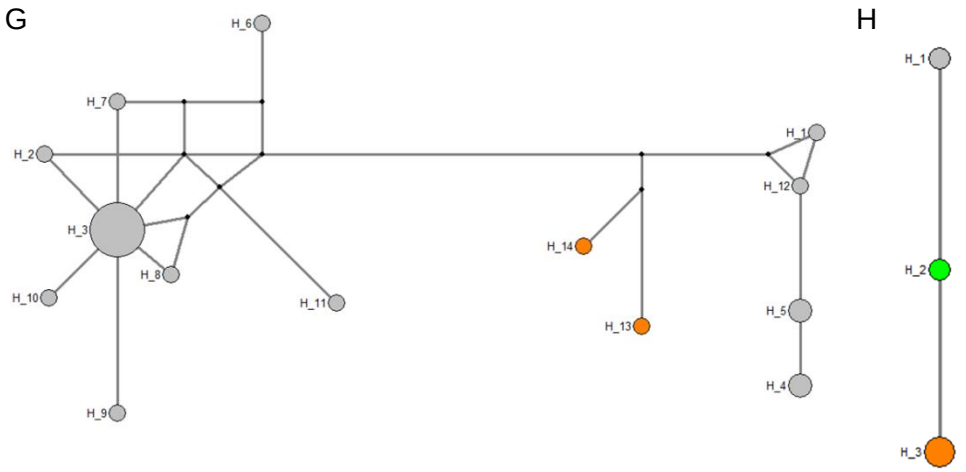


Figure 4 (continued from previous pages). Phylogenetic networks for genera. A) *Cricotopus*, B) *Microtendipes*, C) *Orthocladius*, D) *Paratanytarsus*, E) *Paratendipes*, F) *Phaenopsectra*, G) *Rheocricotopus*, H) *Tvetenia*. Grey for GenBank haplotypes, and Blue, Green and Orange for Bingley Island (BI), Reims Way (RW) and Westgate Towers (WT), respectively.

The abundance (and percentage) of chironomids varied considerably per site (Figure 5A, B). *Microtendipes* (n = 17) was the most abundant genus in the River Stour and it was found in all sites, but it was the only genus found in BI. *Paratendipes* (n = 4), only found in RW, was the next most abundant genus found in the River Stour, followed by *Tvetenia* and *Cricotopus* (n = 3), *Chironomus*, *Orthocladius* and *Rheocricotopus* (n = 2), and *Paratanytarsus* and *Phaenopsectra* (n = 1). *Cricotopus*, *Paratanytarsus*, *Phaenopsectra* and *Rheocricotopus* were only found in WT.

In terms of genus richness per site, WT was site with the highest number of genera, with eight out of the nine genera found in this study. Five genera were collected from RW, including *Microtendipes*, *Tvetenia*, *Chironomus*, *Orthocladius* and *Paratendipes*. As mentioned above, BI was the least genus-rich site with only one genus recorded.

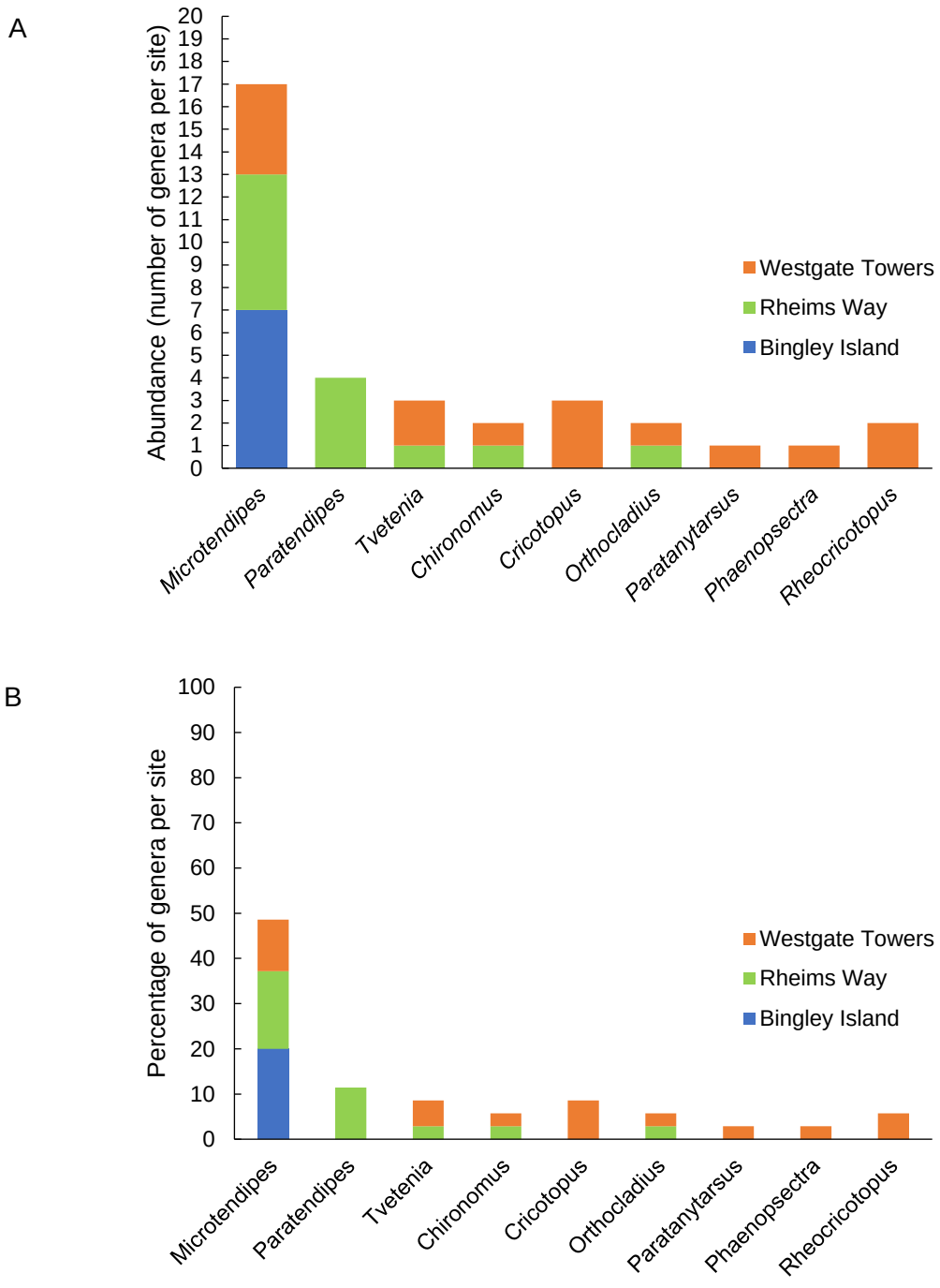


Figure 5. A) Number (abundance) and B) percentage of chironomid genera for each collection site in the River Stour.

Simpson's diversity index for the River Stour was D_S = 0.49. There were substantial differences among sites: BI had a total of seven mtDNA sequences but all belonging to one

genus resulting in D_S = 0. RW had 13 mtDNA sequences from five genera resulting in D_S = 0.67. WT had 17 mtDNA sequences from eight genera resulting in D_S = 0.82.

The Bray-Curtis similarity analysis showed that RW and BI were more similar compared to WT (Fig. 6), although RW and WT were in the same course of the river. However, pairwise Bray-Curtis similarity was not very high, with about 60% between RW and BI, and 40% between WT and RW + BI, and an average Bray-Curtis similarity = 38%.

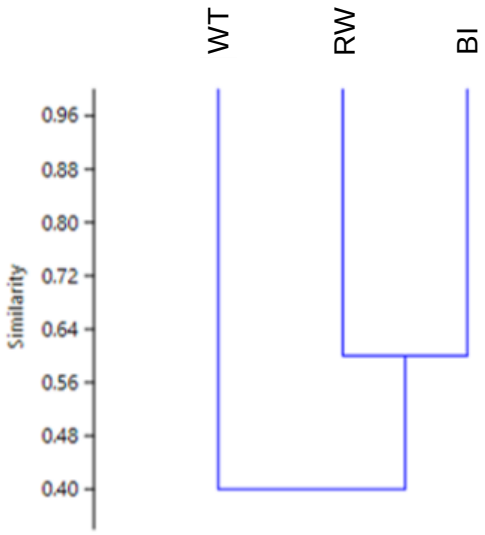


Figure 6. Bray-Curtis similarity among collection sites. BI = Bingley Island, RW = Rheims Way, WT = Westgate Towers.

Discussion

DNA barcoding chironomids

DNA barcoding, the identification of an organism based on molecular data (Hebert, Cywinska and Ball 2003), is an important tool in many areas within ecology, evolution, systematics and conservation (Schindel and Miller 2005). Since its inception, it has been possible to describe and understand the biodiversity of organisms that are difficult to identification using morphological characteristics or because of high biological diversity (Schindel and Miller 2005, Kress *et al.* 2015).

DNA barcoding has simplified the discovery of cryptic species formerly unrecognized through the analysis of morphological features alone (Kress *et al.* 2015). The study of non-biting midges – chironomids – has benefited from the use of this molecular tool, a relevant area in ecology due to the roll of this insect in the ecosystem and as bioindicator species (Ruse and Wilson 1994).

In this study, the first aim was to identify non-biting midge larvae from the River Stour in Canterbury, Kent to the taxonomic level of genus and/or species through the use of DNA barcoding. The mtDNA gene CO1 was chosen as it has been shown to be variable enough to distinguish within the genus level (Puillandre *et al.* 2012). This research project has shown that DNA barcoding is a useful and cost-effective method for the putative identification of chironomid midges. The results showed that DNA barcoding of chironomids using CO1 worked as a quick and reliable identification tool of non-biting midges without the need for morphological identification, as previously shown elsewhere (Failla *et al.* 2015).

Using BLAST, it was possible to obtain a putative identity to the level of genus for the chironomids from the River Stour without the need of relying on morphological characters. Although BLAST only gives a tentative (or putative) genus name to the sample, it provides a quantitative value based on percent identity of the query sequence. This identity value can then be used to detect the best match (Wilson *et al.* 2011). This does not mean that the

sequence belongs to the genus, but a high percent identity gives confidence to the researcher until a complete morphological study is completed. However, when DNA barcoding is used to generate a library – combining morphological identification and DNA sequences of voucher specimens – a robust identification of further samples from the same region is possible (Stepanović *et al.* 2016).

In this study, a total of 33 samples were amplified successfully and DNA-barcoded to GenBank data. The percent identity obtained for each sample was below 100% but, based on the joint results of the BLAST search (% Identity and E value on first and second hits) and the phylogenetic tree, it is likely that the 'putative' genus-level identification was adequate. Despite this, the morphological identification of the chironomid larvae from the River Stour would still be required to confirm the specific identity of the samples. For example, for species-level identification of species of *Tanytarsus* (non-biting midges), a 4-5 % threshold was considered appropriate (Lin, Stur and Ekrem 2015). Therefore, a minimum of 90% identity between query and subject mtDNA sequences, the low E values, and the equivalent first and second hits obtained here could be considered sufficient to attribute a reliable putative genus identity and perform a genus-level community structure analysis of chironomids.

To find out why the identity of the mtDNA sequences obtained here was not 100%, the source information from GenBank was obtained (see Appendix Table S1). It is likely that there was a geographical effect rather than an analytical issue; for example, the sequences identified as *Paratendipes* (first hit BLAST Accession Number: KT707583) had the lowest percentage identity at 89%, but this genus has an almost global distribution (Ashe, Murray and Reiss 1987). This GenBank file (KT707583) was collected on 10th July 2015 in the *rare* Charitable Research Reserve in Ontario, Canada (Telfer *et al.* 2015), thousands of kilometers away from the River Stour. The sequences identified as *Cricotopus* (first hit BLAST Accession Number: KC250771), *Phaenopsectra* (first hit BLAST Accession Number: KC250831) and *Tvetenia* (first hit BLAST Accession Number: KT248931) all had the highest percentage identity at 99%, but all samples in GenBank came from Europe. KC250771 was collected on 30th May 2010 in the Åland Islands on the Baltic Sea Coast, Finland (Brodin *et al.* 2013),

KC250831 was collected on 9th September 2010 in Södermanland County on the Baltic Sea Coast, Sweden (Brodin *et al.* 2013), and KT248931 was collected on 12th September 2014 in Trondheim, Sør-Trøndelag, Norway (Kranzfelder, Ekrem and Stur 2016). Furthermore, in Table 1, all collection sites had an average percentage identity over 90%: Westgate Towers (WT) had the highest average percentage identity at 95.6%, the lowest average percentage identity was for Rheims Way (RW) at 92.5%. WT probably had a higher average percentage identity compared to the other sites because the diversity of genera had also been studied in Europe. The samples in this study that were similar to the Canadian samples could still be considered congeneric, albeit from different geographical populations and possibly belonging to different species; however, phylogeographic studies on each genus are needed to elucidate this.

From the results, it is evident that the geographical origin of the subject data from GenBank will affect the percent identity. This means that for adequate DNA barcoding, a preliminary 'library' and voucher specimens (identified morphologically and through DNA barcoding) from the study region are still needed. In Kent, this is probably the first study of this kind; therefore, there was no library or voucher specimens available in GenBank to match the sample with. Nonetheless, the 'putative' genus data can be used with caution. In a study by Weeraratne, Surendran and Karunaratne (2018), the authors used morphological and genetic identification of mosquitoes and found that genetic characterisation can be successfully used for mosquito species identification; the authors also showed that when traditional morphological identification failed (due to damaged specimens and/or indistinguishable characteristics), DNA barcoding could accurately identify the sample to the species level.

To produce a more complete picture of the chironomid diversity in the River Stour, more samples are needed from more sites and throughout a longer period of time. In this study, there was only a 35.5% success rate in the DNA sequencing of the chironomids; therefore, more samples are needed per site to obtain a representative diversity index of the River Stour and to infer the community structure. It is unknown why several samples failed to amplify; however, this is not uncommon in molecular studies, and there can be many reasons

for amplification failure including: contamination of DNA samples, problems within purifying DNA samples and incorrect storage and transportation of samples (Sundquist and Bessetti 2005, Arbeli and Fuentes 2007, Straube and Juen 2013). Also, as the head capsule and final segment were removed, it is possible that the smallest larval specimens were left with very low amounts of tissue for DNA extraction, possibly resulting in low DNA concentration for PCR. Future trials with different quantities of DNA extraction or primer concentrations could result in positive PCRs. The DNA has been stored and it is available for future work and troubleshooting.

The phylogenetic tree was used as a complementary step to BLAST and in helping with the assignment of putative names to each sample based. The phylogenetic tree showed the evolutionary relationships among the samples from the River Stour and all other chironomids from GenBank. The bootstrap support for the higher nodes of the Neighbour-Joining tree was high showing that the relationships between River Stour samples and GenBank samples can be given with confidence. The lack of bootstrap support for basal nodes was not surprising given the difficulties encountered in phylogenetic studies of insects (Letsch *et al.* 2012); however, as mentioned above, solving the phylogenetic relationships of chironomids was beyond the aims of the study. Here, the percent identities from BLAST search and the phylogenetic tree using CO1 data from GenBank can be used together to provide confidence in the DNA barcoding results. Nonetheless, in molecular systematics, obtaining a good geographic coverage is important while reconstructing intra- or inter-specific phylogenies, and the mtDNA data generated here can be useful for future phylogenetic studies on this diverse family of flies.

Based on the phylogenetic tree, several of the genera appeared to be paraphyletic, including *Rheocricotopus*, *Thienemanniella*, *Corynoneura*, *Limnophyes*, *Metriocnemus*, *Micropsectra*, *Paratanytarsus*, and *Tanytarsus*. This could be the result of using a Neighbour-Joining tree (a genetic distance tree) rather that a more robust phylogenetic reconstruction like Bayesian phylogenetics that uses more complex evolutionary models and tree searching methods (Huelsenbeck, Ronquist and Hall 2001). However, with the growing amount of

mtDNA data available in GenBank and the coverage of the generic diversity of Chironomidae from samples from all around the world, including now South East England, it might be interesting to explore the molecular systematics of this group of insects. There are not many studies on chironomid systematics, but two studies can be highlighted: In a molecular systematics study, the phylogenetic structure of midges of the subfamily Chironominae was inferred from the amino acid sequence of CO1 (Demin, Polukonova and Mugue 2011), confirming the subdivision of Chironominae into two tribes, Chironomini and Tanytarsini, and that *Micropsectra* is formed by a large polyphyletic cluster that includes *Virgotanytharsus*, *Rheotanytarsus*, *Kenopsectra*, and *Parapsectra*. Another study explored the evolutionary history of two species of Orthocladiinae (Diptera: Chironomidae) from Lake Baikal (Eastern Siberia) (Kravtsova *et al.* 2014). The authors used CO1 because this marker provides good phylogenetic signal at the lower inter-species level (with sequence divergences up to 13%) and does not exhibit saturation of transitions and transversions (Kravtsova *et al.* 2014).

The genetic diversity of chironomids depended on the number of samples found per site, with BI having the lowest diversity values overall and only eight sequences. The comparison with GenBank data allowed putting the River Stour sites into contrast with a global data set, and there appears to be high genetic diversity of chironomids in the River Stour, even in a small stretch of river. Although the GenBank data set was about 15 times larger than the River Stour data set, in terms of haplotype diversity the results were very similar. This was because in the River Stour there were mostly different haplotypes across the three sites. Perhaps it is more important to look at the nucleotide diversity. The GenBank data set had much higher nucleotide diversity, but in comparison to other studies on chironomids, the genetic diversity ion the River Stour was also high. For example, in two steams in La Selva Biological Station, Costa Rica chironomids had haplotype diversity values of 0.398 and 0.261 and nucleotide diversity values of 0.005 and 0.008 based on 129 specimens from four species (Small, Wares and Pringle 2011). Two populations of *Sergentia baicalensis* (n = 65) in Lake Baikal, Siberia, Russia had haplotype diversity values of 0.900 and nucleotide diversity values of 0.008 (Kravtsova *et al.* 2015).

Community structure of chironomids in the River Stour

The non-biting midge family Chironomidae is divided into 11 subfamilies: Aphroteniinae, Buchonomyiinae, Chilenomyiinae, Chironominae, Diamesinae, Orthocladiinae, Podonominae, Tanypodinae, Telmatogetoninae and Usambaromyiinae (Leung, Pinder and Edward 2011, Epler 2001, Ashe and O'Connor 2009). From these, only two subfamilies have been identified in the River Stour: Chironominae and Orthocladiinae. The genera *Microtendipes*, *Paratendipes*, *Paratanytarsus*, *Chironomus* and *Phaenopsectra* belong to the subfamily Chironominae. The genera *Cricotopus*, *Tvetenia*, *Rheocricotopus* and *Orthocladius* belong to the subfamily Orthocladiinae.

Based on the taxonomic classification, it appears that the sites in the River Stour were very different in the generic diversity and composition. BI only had one genus, while WT and RW showed eight and five and a different composition of chironomids. This could be due to low sample size per site as well as chance. It was surprising that only one genus was detected in BI, since this is an arm of the river that appears to have different microenvironmental conditions (Supervisor's and Richard Vane-Wright's personal observations). In this particular case, the low generic (taxonomic) diversity in BI impacted the genetic diversity for this site. It was also surprising that BI and RW were more similar to each other than to WT in terms on Bray-Curtis similarity, because WT and RW were geographically closer. The higher Bray-Curtis similarity between BI and RW was due to the few genera detected in each site, while WT showed genera not found elsewhere, namely *Rheocricotopus*, *Paratanytarsus*, *Phaenopsectra* and *Cricotopus*. WT had the highest richness and abundance of chironomids, also resulting in the higher Simpson's diversity index than the other two sites. It would be expected that further sampling and in different times of the year may show greater richness and diversity of chironomids.

There is no other study available on chironomids on the River Stour to contrast the results shown here, so this work represents the first attempt to characterise the community of chironomids. Also, there are not many studies that estimate richness or diversity indexes

based on the DNA barcodes (and putative generic or species names). However, in a study at two river sites in White Clay Creek, Pennsylvania, USA, Sweeney *et al.* (2011) showed that Simpson's diversity indexes were D_S = 0.91 and 0.92 (n = 1579 freshwater invertebrates, from which 45% were Chironomidae), and richness per site was 45 and 70 Chironomidae species. Another study on freshwater invertebrates reporting Simpson's diversity index, but based on morphological identification, showed values ranging from 0.86 to 0.95 in Matapedia River, an Atlantic salmon river of the Gaspesie peninsula (Québec, Canada) (Gillis and Chalifour 2010). The diversity values in White Clay Creek and Matapedia River were considered to be high, while in the River Stour the diversity (overall) was intermediate, but the total number of samples analysed in those studies were also much higher and spanned a longer stretch of the rivers.

The phylogenetic networks per genus reflected the Bray-Curtis similarity results. However, one important pattern that emerged was that all haplotypes from the River Stour were genetically different from the GenBank data. Once again, this was probably due to the geographic origin of the GenBank data. Although the phylogenetic networks were very limited in terms of sample sizes, they reflect that the haplotypes in the River Stour are different from others in GenBank. Only with a more in-depth intra-specific study it would be possible to generate a clear phylogeographic view of haplotype diversity. A phylogeographic study on chironomids from northeast Queensland, Australia using CO1 phylogenetic networks revealed additional shared and unique haplotypes compared with an earlier study, and found a starburst radiation pattern of haplotypes from one common haplotype, an indication that it may be the ancestral haplotype from which other haplotypes derived from and spread geographically (Krosch *et al.* 2011).

DNA barcodes of stream macroinvertebrates can improve descriptions of community structure by increasing the total number of taxa identified by nearly 40% and showing a marked difference between identifications made by experts vs. DNA barcoding (Sweeney *et al.* 2011). This difference between identification by experts and DNA barcode was greatest for

Chironomidae, the most abundant and diverse macroinvertebrate family in a river in White Clay Creek, Pennsylvania, USA (Sweeney *et al.* 2011).

Future directions of DNA barcoding in the River Stour

Based on the mtDNA diversity found here with a limited sample size, it can be assumed that there is a high and cryptic genetic diversity at the level of genus in chironomids in the River Stour which is worth exploring further.

Within each genus, it is possible that there is phylogeographic and population genetic structure, even along the same river (Kébé *et al.* 2017). Further molecular work could be carried out in other sites to characterise the full community structure present in the River Stour. In addition, continued sample collections could be done to monitor the biodiversity of non-biting midges and how the generic composition changes over periods of time.

There are still broad knowledge gaps in chironomid species in the United Kingdom and continued research is needed to fill the gap. DNA barcoding, as shown in this study, can be used reliably for rapid identification to the taxonomic level of genus and/or species. GenBank can be used as a data base to obtain temporary putative identifications until a more careful identification based on morphological characteristics is able to be performed.

DNA barcoding has advantages and limitations. Among several advantages, DNA barcoding is suitable for all life stages of a species, it is possible to differentiate species with little phenotypic differences, and small amounts of biological material is needed (Dudu *et al.* 2016). Some of the limitations of DNA barcoding are that the DNA barcodes are only assigned a putative name, which still needs to be verified by experts, it depends on having a good 'library' available for the taxonomic group of interest and for the geographical region being sampled, and it currently relies on the understanding of mitochondrial inheritance of a single molecular marker (Frezal and Lebois 2008).

Further considerations of DNA-based identification

DNA barcoding is now shifting to the use of high throughput sequencing technique and environmental DNA (eDNA) (Günther *et al.* 2018) and might render traditional DNA barcoding obsolete. This new approach to determining taxonomic diversity is called metabarcoding, which means the parallel sequencing of all DNA extracted from environmental samples (mostly water samples), resulting in millions of DNA sequences of any organism that has left any trace of DNA in the water sample (Deiner *et al.* 2017). An advantage of eDNA and metabarcoding is the potential characterization of different regions of the DNA while using different primers targeting mtDNA and nuclear DNA, and also the amount of nucleotide data that can be retrieved from small amounts of environmental samples without the need to collect the actual specimen.

Although the techniques are becoming available and affordable for carrying out ecological studies, they still require very specialized equipment that is not available in most molecular biology labs, and they also need complex bioinformatic pipelines for the analysis of the genetic data. Just like DNA barcoding, the correct identification of species will also depend on very complete libraries based on the morphological identifications of the species found on the sites, or on using statistical methods for delimiting Operational Taxonomic Units (OTUs).

At this stage, traditional DNA barcoding is still useful and will probably continue to be part of the molecular tool kit for assessing richness and diversity, and monitoring community structure.

Conclusions

This study had two main aims, 1) to identify non-biting midges in the River Stour using CO1 DNA barcoding, and 2) to compare the genetic composition at various sites along the river and characterise the community structure of non-biting midges based on the DNA barcoding data. Here, the CO1 sequences were submitted to GenBank using BLAST and putative generic names were obtained. Although the total number of samples was lower than expected (only 33 DNA sequences were obtained) and there several failed amplifications, it was possible to identify chironomids in the River Stour. With the generic classification, it was also possible to infer the community structure across three sites and to compare the genetic diversity. Based on the literature, there are not many DNA barcoding studies that report richness, diversity and community structure resulting from the DNA barcodes. Most studies discussed here mostly deal with the process of DNA barcoding and species delimitation rather than the ecological implications of species diversity. This makes the study in the River Stour valuable, since it adds to the still limited literature on chironomid diversity.

The DNA sequencing of chironomid larvae in this project was useful for providing a first picture of the community structure in the River Stour. There were nine genera in the River Stour, and different abundances across the three sites. The DNA-based approach allowed to assign with confidence each mtDNA sequence a putative name, without the need of a time-consuming and expert-based morphological approach. Although the identity results were based on best match in GenBank and confirmed with phylogenetic methods, the results are reliable in that both methods returned complementary and consistent information. There is no method that is 100% reliable, and the best was to assess the community structure using freshwater invertebrates would always be a combination of morphological and DNA-based approaches; however, for a quick assessment of the river diversity and community structure, the DNA-based approach is quicker. Notably, no sample collected had a 100% identity when entered as a BLAST search in the NCBI GenBank data base, but the highest percentage given

was 99% identity and the lowest 89% identity. The percent identity results could be attributed to the geographical origin of the GenBank data and the lack of more regional studies in the UK or Kent. For accurate identification of Chironomidae and other insects through DNA barcoding, the NCBI GenBank data base needs to be updated with more chironomid DNA sequences to facilitate future research.

The DNA sequences from the River Stour can be submitted to GenBank to be added to the data base to allow for the assistance with future research in this area. The DNA sequences (haplotypes) were all different from those from GenBank, which indicates that there is potentially much diversity in the River Stour worth exploring. However, to make a more valuable contribution, full taxonomic identification using morphological characteristics of the samples obtained in this study needs to be carried out. Additional research, continuing from this project, would be to observe the morphology of the head and posterior segments of the chironomids that have been sequenced, classify the sample, and verify the putative genera of the DNA sequences. The head and posterior segments could be used for the morphological identification and for generating a 'library' and voucher specimens from the River Stour. The samples are available at CCCU.

Future work within the same topic would be creating and maintaining a monitoring schedule for chironomid genera/species in the River Stour and study how the composition of genera/species changes over time, between the collection sites. For this, the regular sampling of invertebrates is needed, and resources need to be committed to this aim. It is important to consider that in any biomonitoring project the time of year in which samples are collected can change the presence and abundance of invertebrates in the River Stour. Careful consideration has to go into this, and the proper study of the community structure in the river has to involve regular sampling.

Although with limitations, the rapid method of identification used here using DNA barcoding has provided the first insight into the larval community structure of non-biting midges in the River Stour, and it can be used as a baseline for future studies in the region and in comparative studies carried out elsewhere.

Acknowledgments

I would like to thank my supervisor Dr Rodrigo Vega for his patience, guidance and support throughout this research project. Professor Dick Vane-Wright provided an initial idea and formulation of the research project using DNA barcoding. I would also like to thank the two examiners, Dr Emilia Bertolo-Pardo (CCCU) and Dr Nick Littlewood (University of Cambridge) who provided a careful review of this thesis, as well as the Life Sciences laboratory technicians at Canterbury Christ Church University for their help in the laboratory during this time. Permission for access to the sites in the River Stour was granted to Dr Rodrigo Vega and Professor Dick Vane-Wright by the Westgate Parks, and I would like to thank them for allowing the river sampling for carrying out this project.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J., 1990. Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), pp.403-410.

Arbeli, Z. and Fuentes, C.L., 2007. Improved purification and PCR amplification of DNA from environmental samples. *FEMS Microbiology Letters*, *272*(2), pp.269-275.

Ashe, P., Murray, D.A. and Reiss, F., 1987. The zoogeographical distribution of Chironomidae (Insecta: Diptera). *Annals of Limnology*, 23(1), pp.27-60.

Ashe, P. and O'Connor, J.P., 2009. *A world catalogue of Chironomidae (Diptera), Part 1: Buchonomyiinae, Chilenomyiinae, Podonominae, Aphroteniinae, Tanypodinae, Usambaromyiinae, Diamesinae, Prodiamesinae and Telmatogetoninae*. *Irish Biogeographical Society*.

Bailey, K.D., 1994. *Typologies and taxonomies: an introduction to classification techniques* (Vol. 102). Sage.

Barcaccia, G., Lucchin, M. and Cassandro, M., 2015. DNA barcoding as a molecular tool to track down mislabeling and food piracy. *Diversity*, *8*(1), p.2.

Barrett, R.D. and Hebert, P.D., 2005. Identifying spiders through DNA barcodes. *Canadian Journal of Zoology*, *83*(3), pp.481-491.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Sayers, E.W., 2008. GenBank. *Nucleic Acids Research*, *37*(suppl_1), pp.D26-D31.

Beutel, R.G. and Pohl, H., 2006. Endopterygote systematics–where do we stand and what is the goal (Hexapoda, Arthropoda)? *Systematic Entomology*, *31*(2), pp.202-219.

Bray, J.R. and Curtis, J.T., 1957. An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, *27*(4), pp.325-349.

Brodin, Y., Ejdung, G., Strandberg, J. and Lyrholm, T., 2013. Improving environmental and biodiversity monitoring in the Baltic Sea using DNA barcoding of Chironomidae (Diptera). *Molecular Ecology Resources*, *13*(6), pp.996-1004.

Brooks, S.J. and Langdon, P.G., 2014. Summer temperature gradients in northwest Europe during the Late glacial to early Holocene transition (15–8 ka BP) inferred from chironomid assemblages. *Quaternary International*, *341*, pp.80-90.

Cameron, S.L., 2014. Insect mitochondrial genomics: implications for evolution and phylogeny. *Annual Review of Entomology*, *59*, pp.95-117.

Carvalho, C.J.B.D. and Mello-Patiu, C.A.D., 2008. Key to the adults of the most common forensic species of Diptera in South America. *Revista Brasileira de Entomologia*, *52*(3), pp.390-406.

Casey, R.M., 2005. BLAST sequences aid in genomics and proteomics. *Business Intelligence Network*.

Chang, J.C., Shulmeister, J., Woodward, C., Steinberger, L., Tibby, J. and Barr, C., 2015. A chironomid-inferred summer temperature reconstruction from subtropical Australia during the last glacial maximum (LGM) and the last deglaciation. *Quaternary Science Reviews*, *122*, pp.282-292.

Chase, M.W., Soltis, D.E., Olmstead, R.G., Morgan, D., Les, D.H., Mishler, B.D., Duvall, M.R., Price, R.A., Hills, H.G., Qiu, Y.L. and Kron, K.A., 1993. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene rbcL. *Annals of the Missouri Botanical Garden*, pp.528-580.

Chase, M.W. and Fay, M.F., 2009. Barcoding of plants and fungi. *Science*, *325*(5941), pp.682-683.

Chien, A., Edgar, D.B. and Trela, J.M., 1976. Deoxyribonucleic acid polymerase from the extreme thermophile Thermus aquaticus. *Journal of Bacteriology*, *127*(3), pp.1550-1557.

Clary, D.O. and Wolstenholme, D.R., 1985. The mitochondrial DNA molecule of *Drosophila yakuba*: nucleotide sequence, gene organization, and genetic code. *Journal of Molecular Evolution*, *22*(3), pp.252-271.

Coe, R.L., Freeman, P. and Mattingly, P.F., 1953. Handbooks for the identification of British insects, Vol. 9, 2, Diptera, Nematocera.-R. In *Ent. Soc. Lond* (pp. 121-206).

Coffman, W.P. and Ferrington Jr, L.C., 1996. *Chironomidae. An introduction to the aquatic insects of North America*, *3*, pp.635-754.

Cranston, P. (2008), Keys to The Adult Male Chironomidae of Britain and Ireland - Edited by P.H. Langton and L.C.V. Pinder. *A Guide to The Identification of Genera of Chironomid Pupal Exuviae Occurring in Britain and Ireland (Including Common Genera From Northern Europe) and Their Use In Monitoring Lotic and Lentic Fresh Waters* - Edited by R.S. Wilson and L.P. Ruse. *Systematic Entomology*, 33: 215-217.

Dayrat, B., 2005. Towards integrative taxonomy. *Biological Journal of the Linnean Society*, *85*(3), pp.407-417.

De Jong, M.A., Wahlberg, N., Van Eijk, M., Brakefield, P.M. and Zwaan, B.J., 2011. Mitochondrial DNA signature for range-wide populations of Bicyclus anynana suggests a rapid expansion from recent refugia. PloS One, 6(6), p.e21385.

De Mandal, S., Chhakchhuak, L., Gurusubramanian, G. and Kumar, N.S., 2014. Mitochondrial markers for identification and phylogenetic studies in insects–A Review. *DNA Barcodes*, *2*(1).

Deiner, K., Bik, H.M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D.M., De Vere, N. and Pfrender, M.E., 2017. Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, *26*(21), pp.5872-5895.

Demin, A.G., Polukonova, N.V. and Mugue, N.S., 2011. Molecular phylogeny and the time of divergence of midges (Chironomidae, Nematocera, Diptera) inferred from a partial nucleotide sequence of the cytochrome oxidase I gene (COI). *Russian Journal of Genetics*, 47(1168), https://doi.org/10.1134/S1022795411100061

Dewey, H., Wooldridge, S.W., Cornes, H.W. and Brown, E.E.S., 1925. The geology of the Canterbury District: With special reference to the Whitsuntide Excursion, 1925. *Proceedings of the Geologists' Association*, *36*(3), pp.257-IN11.

Dobson, M., Pawley, S., Fletcher, M. and Powell, A., 2012. *Guide to freshwater invertebrates.* Freshwater Biological Association.

Eggermont, H. and Verschuren, D., 2007. Taxonomy and diversity of Afroalpine Chironomidae (Insecta: Diptera) on Mount Kenya and the Rwenzori mountains, East Africa. *Journal of Biogeography*, *34*(1), pp.69-89.

Ekrem, T., Stur, E. and Hebert, P.D., 2010. Females do count: Documenting Chironomidae (Diptera) species diversity using DNA barcoding. *Organisms Diversity & Evolution*, *10*(5), pp.397-408.

Environment Agency. 2013. Stour Abstraction licensing strategy.

Environment Agency. 2014. The Stour Management Catchment. A summary of information about the water environment in the Stour management catchment.

Environment Agency. 2016. DATASHEET: Nitrate Vulnerable Zone (NVZ) designation 2017 - Surface Water.

Epler, John H. "Identification manual for the larval Chironomidae (Diptera) of North and South Carolina." *A guide to the taxonomy of the midges of the southeastern United States, including Florida. Special Publication SJ2001-SP13. North Carolina Department of Environment and Natural Resources, Raleigh, NC, and St. Johns River Water Management District, Palatka, FL* 526 (2001).

Failla, A.J., Vasquez, A.A., Fujimoto, M. and Ram, J.L., 2015. The ecological, economic and public health impacts of nuisance chironomids and their potential as aquatic invaders. *Aquatic Invasions*, *10*(1).

Failla, A.J., Vasquez, A.A., Hudson, P., Fujimoto, M. and Ram, J.L., 2016. Morphological identification and COI barcodes of adult flies help determine species identities of chironomid larvae (Diptera, Chironomidae). *Bulletin of Entomological Research*, *106*(1), pp.34-46.

Farrant, A.R. and Aldiss, D.T., 2002. A geological model of the North Downs of Kent: the River Medway to the River Great Stour.

Ferrington, L.C., 2008. Global diversity of non-biting midges (Chironomidae; Insecta-Diptera) in freshwater. *Hydrobiologia*, *595*(1), p.447.

Floyd, R.M., Wilson, J.J. and Hebert, P.D., 2009. DNA barcodes and insect biodiversity. *Insect Biodiversity: Science and Society. Oxford: Blackwell Publishing*, pp.417-431.

Fluxus Technology Ltd. 2015. Network 5.0.0.0 User Guide.

Folmer, O., Black, M., Hoeh, W., Lutz, R., Vrijenhoek, R., 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, *3*(5), pp.294-299.

Fortin, M.C., Medeiros, A.S., Gajewski, K., Barley, E.M., Larocque-Tobler, I., Porinchu, D.F. and Wilson, S.E., 2015. Chironomid-environment relations in northern North America. *Journal of Paleolimnology*, *54*(2-3), pp.223-237.

Frézal, L. and Leblois, R., 2008. Four years of DNA barcoding: current advances and prospects. *Infection, Genetics and Evolution*, *8*(5), pp.727-736.

Gillis, C.A. and Chalifour, M., 2010. Changes in the macrobenthic community structure following the introduction of the invasive algae *Didymosphenia geminata* in the Matapedia River (Québec, Canada). *Hydrobiologia*, *647*(1), pp.63-70.

Günther, B., Knebelsberger, T., Neumann, H., Laakmann, S. and Arbizu, P.M., 2018. Metabarcoding of marine environmental DNA based on mitochondrial and nuclear genes. *Scientific Reports*, *8*(1), p.14822.

Hajibabaei, M., Janzen, D.H., Burns, J.M., Hallwachs, W. and Hebert, P.D., 2006. DNA barcodes distinguish species of tropical Lepidoptera. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(4), pp.968-971.

Hajibabaei, M., Singer, G.A., Clare, E.L. and Hebert, P.D., 2007. Design and applicability of DNA arrays and DNA barcodes in biodiversity monitoring. *BMC Biology*, *5*(1), p.24.

Hall, T.A., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, 41, pp.95-98.

Hammer, Ř., Harper, D.A.T. and Ryan, P.D., 2001. PAST: Paleontological Statistics Software Package for Education and Data Analysis–Palaeontol. Electron. 4: 9pp.

Hawthorne, P., Wang-Holmes, J., Cawdell-Smith, J. and Trezise, A.E., 2009. DNA Profiling of horse urine samples to confirm donor identity. *Rural Industries Research and Development Corporation*.

Hebert, P.D., Cywinska, A. and Ball, S.L., 2003. Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London B: Biological Sciences*, *270*(1512), pp.313-321.

Hebert, P.D., Stoeckle, M.Y., Zemlak, T.S. and Francis, C.M., 2004. Identification of birds through DNA barcodes. *PLoS Biology*, *2*(10), p.e312.

Hua, G.J., Hung, C.L., Lin, C.Y., Wu, F.C., Chan, Y.W. and Tang, C.Y., 2017. MGUPGMA: A Fast UPGMA Algorithm With Multiple Graphics Processing Units Using NCCL. *Evolutionary Bioinformatics*, *13*, p.1176934317734220.

Huelsenbeck, J.P., Ronquist, F. and Hall, B., 2001. An introduction to Bayesian inference of phylogeny.

Huson, D.H. and Bryant, D., 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23, pp.254-267.

Ilmonen, J. and Paasivirta, L., 2005. Benthic macrocrustacean and insect assemblages in relation to spring habitat characteristics: patterns in abundance and diversity. *Hydrobiologia*, *533*(1-3), pp.99-113.

Jaric, M., Segal, J., Silva-Herzog, E., Schneper, L., Mathee, K. and Narasimhan, G., 2013, December. Better primer design for metagenomics applications by increasing taxonomic distinguishability. In *BMC Proceedings* (Vol. 7, No. 7, p. S4). BioMed Central.

Jing, Y.U., Jian-Hua, X.U.E. and Shi-Liang, Z.H.O.U., 2011. New universal matK primers for DNA barcoding angiosperms. *Journal of Systematics and Evolution*, *49*(3), pp.176-181.

Joly, S., Davies, T.J., Archambault, A., Bruneau, A., Derry, A., Kembel, S.W., Peres-Neto, P., Vamosi, J. and Wheeler, T.A., 2014. Ecology in the age of DNA barcoding: the

resource, the promise and the challenges ahead. *Molecular Ecology Resources*, *14*(2), pp.221-232.

Kébé, K., Alvarez, N., Tuda, M., Arnqvist, G., Fox, C.W., Sembène, M. and Espíndola, A., 2017. Global phylogeography of the insect pest *Callosobruchus maculatus* (Coleoptera: Bruchinae) relates to the history of its main host, *Vigna unguiculata*. *Journal of Biogeography*, *44*(11), pp.2515-2526.

Kelley, J.L., Peyton, J.T., Fiston-Lavier, A.S., Teets, N.M., Yee, M.C., Johnston, J.S., Bustamante, C.D., Lee, R.E. and Denlinger, D.L., 2014. Compact genome of the Antarctic midge is likely an adaptation to an extreme environment. *Nature Communications*, *5*.

Kranzfelder, P., Ekrem, T. and Stur, E., 2016. Trace DNA from insect skins: a comparison of five extraction protocols and direct PCR on chironomid pupal exuviae. *Molecular Ecology Resources*, *16*(1), pp.353-363.

Kravtsova, L.S., Bukin, Y.S., Peretolchina, T.E. and Sherbakov, D.Y., 2015. Genetic differentiation of populations of Baikal endemic *Sergentia baicalensis* Tshern. (Diptera, Chironomidae). *Russian Journal of Genetics*, *51*(7), pp.707-710.

Kravtsova, L.S., Peretolchina, T.E., Triboy, T.I. and Sherbakov, D.Y., 2014. The evolutionary history of two species of Orthocladiinae (Diptera: Chironomidae) from Lake Baikal (Eastern Siberia). *Aquatic Insects*, 36(3-4), pp.171-185.

Kress, W.J., García-Robledo, C., Uriarte, M. and Erickson, D.L., 2015. DNA barcodes for ecology, evolution, and conservation. *Trends in Ecology & Evolution*, *30*(1), pp.25-35.

Krosch, M.N., Baker, A.M., Mather, P.B. and Cranston, P.S., 2011. Spatial population genetic structure reveals strong natal site fidelity in *Echinocladius martini* (Diptera: Chironomidae) in northeast Queensland, Australia. Freshwater Biology, DOI: 10.1111/j.1365-2427.2010.02571.x

Kumar, S., Stecher, G. and Tamura, K., 2015. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0. *Molecular Biology and Evolution* (submitted).

Letsch, H.O., Meusemann, K., Wipfler, B., Schütte, K., Beutel, R. and Misof, B., 2012. Insect phylogenomics: results, problems and the impact of matrix composition. *Proceedings of the Royal Society of London B: Biological Sciences*, *279*(1741), pp.3282-3290.

Leung, A., Pinder, A. and Edward, D., 2011. Photographic guide and keys to the larvae of Chironomidae (Diptera) of south-west Western Australia.

Lima, J., Floyd, R. and Hanner, R.H., 2008, June. DNA barcoding and its relevance to pests, plants and biological control. In *XI International Symposium on the Processing Tomato 823* (pp. 41-48).

Lin, X., Stur, E., Ekrem, T., 2015. Exploring genetic divergence in a species-rich insect genus using 2790 DNA barcodes. *Plos ONE*, *10*(9), e0138993.

Lunt, D.H., Zhang, D.X., Szymura, J.M. and Hewltt, O.M., 1996. The insect cytochrome oxidase I gene: evolutionary patterns and conserved primers for phylogenetic studies. *Insect Molecular Biology*, *5*(3), pp.153-165.

Luo, A., Zhang, A., Ho, S.Y., Xu, W., Zhang, Y., Shi, W., Cameron, S.L. and Zhu, C., 2011. Potential efficacy of mitochondrial genes for animal DNA barcoding: a case study using eutherian mammals. *BMC Genomics*, *12*(1), p.84.

Massaferro, J., Larocque-Tobler, I., Brooks, S.J., Vandergoes, M., Dieffenbacher-Krall, A. and Moreno, P., 2014. Quantifying climate change in Huelmo mire (Chile, Northwestern Patagonia) during the Last Glacial Termination using a newly developed chironomid-based temperature model. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *399*, pp.214-224.

Matz, M.V. and Nielsen, R., 2005. A likelihood ratio test for species membership based on DNA sequence data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1462), pp.1969-1974.

Meusnier, I., Singer, G.A., Landry, J.F., Hickey, D.A., Hebert, P.D. and Hajibabaei, M., 2008. A universal DNA mini-barcode for biodiversity analysis. *BMC Genomics*, *9*(1), p.214.

Mount, D.W., 2004. Using hidden Markov model to align multiple sequences in: Bioinformatics: Sequence and Genome Analysis.

Mullis, K.B., Erlich, H.A., Arnheim, N., Horn, G.T., Saiki, R.K. and Scharf, S.J., Cetus Corp, 1987. *Process for amplifying, detecting, and/or-cloning nucleic acid sequences*. U.S. Patent 4,683,195.

Mwale, M., Dalton, D.L., Jansen, R., De Bruyn, M., Pietersen, D., Mokgokong, P.S. and Kotzé, A., 2016. Forensic application of DNA barcoding for identification of illegally traded African pangolin scales. *Genome*, *60*(3), pp.272-284.

NCBI National Center for Biotechnology Information. 2019. Our Mission. [ONLINE] Available at: https://www.ncbi.nlm.nih.gov/home/about/mission/. [Accessed 10 June 2019].

Nicacio, G. and Juen, L., 2015. Chironomids as indicators in freshwater ecosystems: an assessment of the literature. Insect Conservation and Diversity, DOI: 10.1111/icad.12123.

Paasivirta, L., 2014. Checklist of the family Chironomidae (Diptera) of Finland. *ZooKeys*, (441), p.63.

Pečnikar, Ž.F. and Buzan, E.V., 2014. 20 years since the introduction of DNA barcoding: from theory to application. *Journal of Applied Genetics*, *55*(1), pp.43-52.

Pérez, L., Lorenschat, J., Massaferro, J., Pailles, C., Sylvestre, F., Hollwedel, W., Brandorff, G.O., Brenner, M., Gerald, I., Lozano, M.D.S. and Scharf, B., 2013. Bioindicators of climate and trophic state in lowland and highland aquatic ecosystems of the Northern Neotropics. *Revista de Biología Tropical*, *61*(2), pp.603-644.

Pfenninger, M., Nowak, C., Kley, C., Steinke, D. and Streit, B., 2007. Utility of DNA taxonomy and barcoding for the inference of larval community structure in morphologically cryptic Chironomus (Diptera) species. *Molecular Ecology*, *16*(9), pp.1957-1968.

Piggott, M.P. and Taylor, A.C., 2003. Remote collection of animal DNA and its applications in conservation management and understanding the population biology of rare and cryptic species. *Wildlife Research*, *30*(1), pp.1-13.

Piggott, M.P., 2005. Effect of sample age and season of collection on the reliability of microsatellite genotyping of faecal DNA. *Wildlife Research*, *31*(5), pp.485-493.

Porinchu, D.F. and MacDonald, G.M., 2003. The use and application of freshwater midges (Chironomidae: Insecta: Diptera) in geographical research. *Progress in Physical Geography*, *27*(3), pp.378-422.

Pratheepa, M., Jalali, S.K., Arokiaraj, R.S., Venkatesan, T., Nagesh, M., Panda, M. and Pattar, S., 2014. Insect Barcode Information System. *Bioinformation*, *10*(2), p.98.

Puillandre, N., Lambert, A., Brouillet, S. and Achaz, G., 2012. ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Molecular Ecology*, *21*(8), pp.1864-1877.

Ram, J.L., Banno, F., Gala, R.R., Gizicki, J.P. and Kashian, D.R., 2014. Estimating sampling effort for early detection of non-indigenous benthic species in the Toledo Harbor Region of Lake Erie. *Management of Biological Invasions*, *5*(3), pp.209-216.

Rogers, P. and Dussart, G., 2004. Investigation into the macro-algae community of Pegwell Bay.

Rozas, J., 2009. DNA sequence polymorphism analysis using DnaSP. In *Bioinformatics for DNA sequence analysis* (pp. 337-350). Humana Press.

Ruse L, Wilson RS .1994. Long-term assessment of water and sediment quality of the River Thames using chironomid pupal skins. In: Cranston P (ed), *Chironomids: From genes to ecosystems*. CSIRO Australia, 482 pp.

Ruse, L.P., 2013. Chironomid (Diptera) species recorded from UK lakes as pupal exuviae. *Journal of Entomological and Acarological Research*, *45*(2), p.13.

Sæther OA, Spies M. 2013. Fauna Europaea: Chironomidae. In: Pape T, Beuk P. (Eds) Fauna Europaea: Diptera, Version 2.6. http://www.faunaeur.org/

Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A. and Arnheim, N., 1985. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, *230*(4732), pp.1350-1354.

Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B. and Erlich, H.A., 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, *239*(4839), pp.487-491.

Saitou, N. and Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution, 4(4), pp.406-425.

Sanger, F., Nicklen, S., and Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), pp.5463–5467.

Savolainen, V., Cowan, R.S., Vogler, A.P., Roderick, G.K. and Lane, R., 2005. Towards writing the encyclopaedia of life: an introduction to DNA barcoding. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *360*(1462), pp.1805-1811.

Schander, C. and Willassen, E., 2005. What can biological barcoding do for marine biology? *Marine Biology Research*, *1*(1), pp.79-83.

Schindel, D.E. and Miller, S.E., 2005. DNA barcoding a useful tool for taxonomists. *Nature*, *435*(7038), pp.17-17.

Schoenbrunner, N.J., Gupta, A.P., Young, K.K. and Will, S.G., 2017. Covalent modification of primers improves PCR amplification specificity and yield. *Biology Methods and Protocols*, *2*(1), p.bpx011.

Seifert, K.A., Samson, R.A., Houbraken, J., Lévesque, C.A., Moncalvo, J.M., Louis-Seize, G. and Hebert, P.D., 2007. Prospects for fungus identification using CO1 DNA barcodes, with Penicillium as a test case. *Proceedings of the National Academy of Sciences*, *104*(10), pp.3901-3906.

Simpson, E.H., 1949. Measurement of diversity. *Nature*, *163*(4148), p.688.

Small, G.E., Wares, J.P. and Pringle, C.M., 2011. Differences in phosphorus demand among detritivorous chironomid larvae reflect intraspecific adaptations to differences in food resource stoichiometry across lowland tropical streams. *Limnology and Oceanography*, *56*(1), pp.268-278.

Stielow, J.B., Lévesque, C.A., Seifert, K.A., Meyer, W., Iriny, L., Smits, D., Renfurm, R., Verkley, G.J.M., Groenewald, M., Chaduli, D. and Lomascolo, A., 2015. One fungus, which genes? Development and assessment of universal primers for potential

secondary fungal DNA barcodes. *Persoonia: Molecular Phylogeny and Evolution of Fungi*, *35*, p.242.

Straube, D. and Juen, A., 2013. Storage and shipping of tissue samples for DNA analyses: A case study on earthworms. *European Journal of Soil Biology*, *57*, pp.13-18.

Stepanović, S., Kosovac, A., Krstić, O., Jović, J. and Toševski, I., 2016. Morphology versus DNA barcoding: two sides of the same coin. A case study of Ceutorhynchus erysimi and C. contractus identification. *Insect Science*, *23*(4), pp.638-648.

Strüder-Kypke, M.C. and Lynn, D.H., 2010. Comparative analysis of the mitochondrial cytochrome c oxidase subunit I (COI) gene in ciliates (Alveolata, Ciliophora) and evaluation of its suitability as a biodiversity marker. *Systematics and Biodiversity*, *8*(1), pp.131-148.

Sundquist, T. and Bessetti, J., 2005. Identifying and preventing DNA contamination in a DNA-Typing Laboratory. *Profiles in DNA. September*, pp.11-13.

Sweeney, B.W., Battle, J.M., Jackson, J.K. and Dapkey, T., 2011. Can DNA barcodes of stream macroinvertebrates improve descriptions of community structure and water quality? *Journal of the North American Benthological Society*, 30(1), pp.195-216.

Tang, H., Song, M.Y., Cho, W.S., Park, Y.S. and Chon, T.S., 2010, January. Species abundance distribution of benthic chironomids and other macroinvertebrates across different levels of pollution in streams. In *Annales de Limnologie-International Journal of Limnology* (Vol. 46, No. 1, pp. 53-66). EDP Sciences.

Theissinger, K., Kästel, A., Elbrecht, V., Leese, F. and Brühl, C., 2018. Using DNA metabarcoding for assessing chironomid diversity and community change in mosquito controlled temporary wetlands. *Metabarcoding and Metagenomics*, *2*, p.e21060.

Telfer, A.C., Young, M.R., Quinn, J., Perez, K., Sobel, C.N., Sones, J.E., Levesque-Beaudin, V., Derbyshire, R., Fernandez-Triana, J., Rougerie, R. and Thevanayagam, A., 2015. Biodiversity inventories in high gear: DNA barcoding facilitates a rapid biotic survey of a temperate nature reserve. *Biodiversity Data Journal*, (3).

ThermoFisher Scientific. 2018. DreamTaq Green PCR Master Mix (2X). https://www.thermofisher.com/order/catalog/product/K1081

Tunnicliffe, V., 1991. The biology of hydrothermal vents: ecology and evolution. *Oceanographic Marine Bio. Ann. Reu.*, *29*, pp.319-407.

Valentini, A., Pompanon, F. and Taberlet, P., 2009. DNA barcoding for ecologists. *Trends in Ecology & Evolution*, *24*(2), pp.110-117.

Vences, M., Nagy, Z.T., Sonet, G. and Verheyen, E., 2012. DNA barcoding amphibians and reptiles. In *DNA Barcodes* (pp. 79-107). Humana Press, Totowa, NJ.

Ward, R.D., Zemlak, T.S., Innes, B.H., Last, P.R. and Hebert, P.D., 2005. DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *360*(1462), pp.1847-1857.

Weeraratne, T.C., Surendran, S.N. and Karunaratne, S.P., 2018. DNA barcoding of morphologically characterized mosquitoes belonging to the subfamily Culicinae from Sri Lanka. *Parasites & Vectors*, *11*(1), p.266.

Wheeler, D. and Bhagwat, M., 2007. BLAST QuickStart: example-driven web-based BLAST tutorial. In *Comparative Genomics Volume 1, Methods in Molecular Biology*, *395-396* (pp.149-76). Humana Press, Totowa, NJ.

Wilkerson, R.C., Linton, Y.M., Fonseca, D.M., Schultz, T.R., Price, D.C. and Strickman, D.A., 2015. Making mosquito taxonomy useful: a stable classification of tribe Aedini that balances utility with current knowledge of evolutionary relationships. *PLoS One*, *10*(7), p.e0133602.

Wilson, J.J., Rougerie, R., Schonfeld, J., Janzen, D.H., Hallwachs, W., Hajibabaei, M., Kitching, I.J., Haxaire, J. and Hebert, P.D., 2011. When species matches are unavailable are DNA barcodes correctly assigned to higher taxa? An assessment using sphingid moths. *BMC Ecology*, *11*(1), p.18.

Witt, J.D., Threloff, D.L. and Hebert, P.D., 2006. DNA barcoding reveals extraordinary cryptic diversity in an amphipod genus: implications for desert spring conservation. *Molecular Ecology*, *15*(10), pp.3073-3082.

Yusseff-Vanegas, S.Z. and Agnarsson, I., 2017. DNA-barcoding of forensically important blow flies (Diptera: Calliphoridae) in the Caribbean Region. *PeerJ*, *5*, p.e3516.

Zhang, A.B., He, L.J., Crozier, R.H., Muster, C. and Zhu, C.D., 2010. Estimating sample sizes for DNA barcoding. *Molecular Phylogenetics and Evolution*, *54*(3), pp.1035-1039.

Glossary

µL: Microlitre. SI measurement of liquids equal to 10^{-6} litres.

µM: Micrometre. SI measurement of length equal to 10^{-6} metres.

% Identity: Percentage Identity. The similarity between two sequences.

Accession: A unique identifier given to a biological polymer sequence.

Base pairs (bp): A pair of complementary bases in a double-stranded nucleic acid molecule.

BioEdit: Software used for sequence alignment editing and analysis.

BLAST: Basic Local Alignment Search Tool. Bioinformatics program for searching DNA sequences.

Bray-Curtis similarity: A statistic used to quantify the similarity between two different sites, based on counts at each site.

Chironomid(ae): Insect family that includes Non-biting midge.

ClustalW: Bioinformatic computer programme used for multiple sequence alignments.

CO1: Cytochrome oxidase 1 also known as Cox1.Mitochondrial DNA.

Diptera: An order of two-winged insects also known as true flies.

DNA-barcoding: Method of species identification using a short section of DNA.

E value: A parameter used to represent the number of matches (hits) that can be expected to be seen by chance.

GenBank: Public online database consisting of nucleotide sequences. Part of the NCBI.

Genus: A taxonomic category that ranks above species and below family.

Haplotype diversity (Hd): The probability that two randomly sampled alleles are different.

Kick Sampling: Sample collection method used for invertebrates in rivers.

Max Score: The highest alignment score of a set of aligned segments from the same subject (data base) sequence.

MtDNA: Mitochondrial DNA.

NCBI: The National Center for Biotechnology Information.

Neighbour-Joining: A clustering method for the creation of phylogenetic trees.

Nucleotide sequences: A sequence of letters (ACGT) that express the order of nucleotides that form alleles within a DNA molecule.

Nucleotide diversity (π): The average number of nucleotide differences per site in pairwise comparisons among DNA sequences.

Phylogenetic network: Any graph used to visualise evolutionary relationships based on networks.

Polymerase Chain Reaction (PCR): A molecular biology method used to make copies of a specific DNA region.

Query Cover: The percentage of a given input sequence (the query) that has aligned with a sequence.

Simpson's diversity index: Calculation used in ecological surveys to calculate the diversity of species and their abundance present in a given area.

Species: A group of living organisms consisting of similar individuals.

Total Score: The sum of alignment scores of all segments from the same subject sequence.

Table S1. GenBank identity (BLAST) of the samples from the River Stour. The sample ID was based on the name of the river (ST = River Stour) and the name of the site (BI = Bingley Island, WT = Westgate Towers, RW = Rheims Way), followed by sample number.

Sample ID	Collection Date	Genus	Species	Max Score	Total Score	Query Cover (%)	E Value	% Identity	BLAST Accession Number
STBI_2	19/10/2016	*Microtendipes*	*pedellus*	915	915	96	< 0.001	93	KC250812
STBI_8	19/10/2016	*Microtendipes*	*pedellus*	933	933	96	< 0.001	93	KC250812
STBI_16	19/10/2016	*Microtendipes*	*pedellus*	941	941	92	< 0.001	94	KC250812
STBI_17	19/10/2016	*Microtendipes*	*pedellus*	905	905	92	< 0.001	92	KC250812
STBI_21	19/10/2016	*Microtendipes*	*pedellus*	929	929	92	< 0.001	93	KC250812
STBI_31	19/10/2016	*Microtendipes*	*pedellus*	939	939	94	< 0.001	94	KC250812
STBI_34	07/04/2017	*Microtendipes*	*pedellus*	946	946	93	< 0.001	94	KC250812
STRW_1	19/10/2016	*Paratendipes*	*N/A*	782	782	97	< 0.001	89	KT707583

(Continued from previous page)

Sample ID	Collection Date	Genus	Species	Max Score	Total Score	Query Cover (%)	E Value	% Identity	BLAST Accession Number
STRW_5	19/10/2016	*Paratendipes*	*N/A*	769	769	93	< 0.001	89	KT707583
STRW_6	19/10/2016	*Paratendipes*	*N/A*	771	771	90	< 0.001	89	KT707583
STRW_7	19/10/2016	*Microtendipes*	*pedellus*	929	929	95	< 0.001	93	KC250812
STRW_8	19/10/2016	*Microtendipes*	*pedellus*	920	920	95	< 0.001	93	KC250812
STRW_9	19/10/2016	*Chironomus*	*riparius*	1040	1040	95	< 0.001	97	HM137928
STRW_10	19/10/2016	*Paratendipes*	*N/A*	761	761	91	< 0.001	89	KT707583
STRW_11	19/10/2016	*Microtendipes*	*pedellus*	857	857	95	< 0.001	92	KC250812
STRW_13	07/04/2017	*Microtendipes*	*pedellus*	935	935	92	< 0.001	94	KC250812

(Continued from previous page)

Sample ID	Collection Date	Genus	Species	Max Score	Total Score	Query Cover (%)	E Value	% Identity	BLAST Accession Number
STRW_19	19/10/2016	*Tvetenia*	*calvescens*	1024	1024	90	< 0.001	99	KT248931
STRW_22	07/04/2017	*Orthocladius*	*N/A*	896	896	94	< 0.001	92	JF287403
STRW_28	07/04/2017	*Microtendipes*	*pedellus*	944	944	93	< 0.001	94	KC250812
STRW_30	07/04/2017	*Microtendipes*	*pedellus*	898	898	95	< 0.001	93	KC250812
STWT_1	19/10/2016	*Cricotopus*	*triannulatus*	981	981	96	< 0.001	95	LC050963
STWT_9	07/04/2017	*Chironomus*	*riparius*	1040	1040	95	< 0.001	97	HM137928
STWT_13	07/04/2017	*Microtendipes*	*pedellus*	922	922	95	< 0.001	93	KC250812
STWT_14	07/04/2017	*Phaenopsectra*	*flavipes*	1122	1122	92	< 0.001	99	KC250831

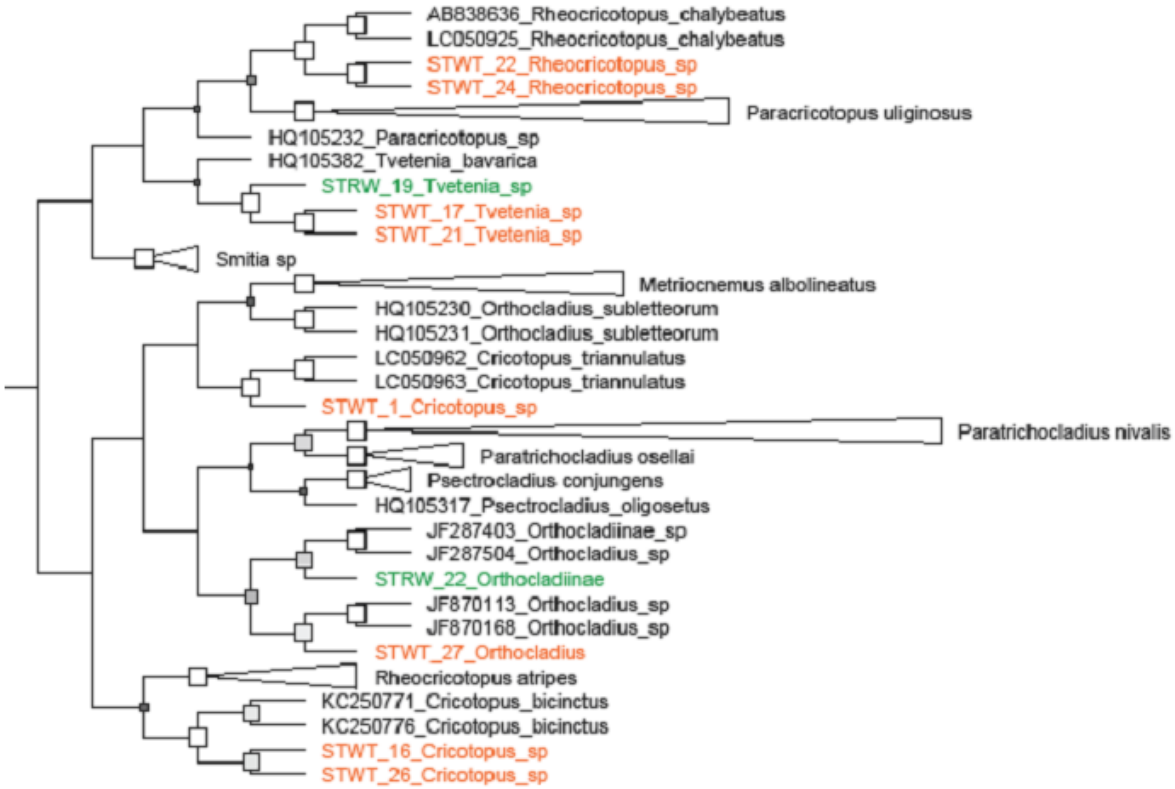
(Continued from previous page)

Sample ID	Collection Date	Genus	Species	Max Score	Total Score	Query Cover (%)	E Value	% Identity	BLAST Accession Number
STWT_15	07/04/2017	*Microtendipes*	*pedellus*	942	942	93	< 0.001	94	KC250812
STWT_16	07/04/2017	*Cricotopus*	*bicinctus*	1120	1120	94	< 0.001	99	KC250771
STWT_17	07/04/2017	*Tvetenia*	*calvescens*	1038	1038	89	< 0.001	98	KT248931
STWT_18	07/04/2017	*Microtendipes*	*pedellus*	935	935	92	< 0.001	94	KC250812
STWT_21	07/04/2017	*Tvetenia*	*calvescens*	1042	1042	86	< 0.001	99	KT248931
STWT_22	19/10/2016	*Microtendipes*	*pedellus*	935	935	95	< 0.001	94	KC250812
STWT_22	07/04/2017	*Rheocricotopus*	*chalybeatus*	896	896	90	< 0.001	91	LC050925
STWT_24	07/04/2017	*Rheocricotopus*	*chalybeatus*	1072	1072	94	< 0.001	97	LC050925

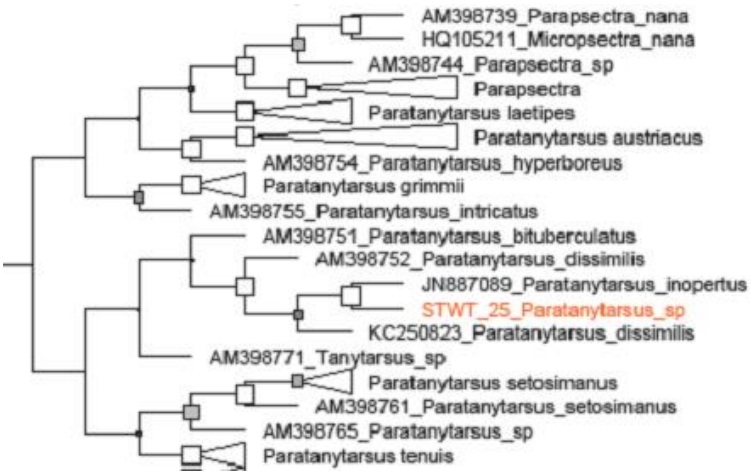
(Continued from previous page)

Sample ID	Collection Date	Genus	Species	Max Score	Total Score	Query Cover (%)	E Value	% Identity	BLAST Accession Number
STWT_25	07/04/2017	*Paratanytarsus*	*inopertus*	935	935	92	< 0.001	93	JN887089
STWT_26	07/04/2017	*Crictopus*	*bicinctus*	1105	1105	92	< 0.001	99	KC250771
STWT_27	07/04/2017	*Orthocladius*	*N/A*	883	883	92	< 0.001	92	JF870168

A


AB838636_Rheocricotopus_chalybeatus
LC050925_Rheocricotopus_chalybeatus
STWT_22_Rheocricotopus_sp
STWT_24_Rheocricotopus_sp
Paracricotopus uliginosus
HQ105232_Paracricotopus_sp
HQ105382_Tvetenia_bavarica
STRW_19_Tvetenia_sp
STWT_17_Tvetenia_sp
STWT_21_Tvetenia_sp
Smittia sp
Metriocnemus albolineatus
HQ105230_Orthocladius_subletteorum
HQ105231_Orthocladius_subletteorum
LC050962_Cricotopus_triannulatus
LC050963_Cricotopus_triannulatus
STWT_1_Cricotopus_sp
Paratrichocladius nivalis
Paratrichocladius osellai
Psectrocladius conjungens
HQ105317_Psectrocladius_oligosetus
JF287403_Orthocladiinae_sp
JF287504_Orthocladius_sp
STRW_22_Orthocladiinae
JF870113_Orthocladius_sp
JF870168_Orthocladius_sp
STWT_27_Orthocladius
Rheocricotopus atripes
KC250771_Cricotopus_bicinctus
KC250776_Cricotopus_bicinctus
STWT_16_Cricotopus_sp
STWT_26_Cricotopus_sp

B


AM398739_Parapsectra_nana
HQ105211_Micropsectra_nana
AM398744_Parapsectra_sp
Parapsectra
Paratanytarsus laetipes
Paratanytarsus austriacus
AM398754_Paratanytarsus_hyperboreus
Paratanytarsus grimmii
AM398755_Paratanytarsus_intricatus
AM398751_Paratanytarsus_bituberculatus
AM398752_Paratanytarsus_dissimilis
JN887089_Paratanytarsus_inopertus
STWT_25_Paratanytarsus_sp
KC250823_Paratanytarsus_dissimilis
AM398771_Tanytarsus_sp
Paratanytarsus setosimanus
AM398761_Paratanytarsus_setosimanus
AM398765_Paratanytarsus_sp
Paratanytarsus tenuis

C

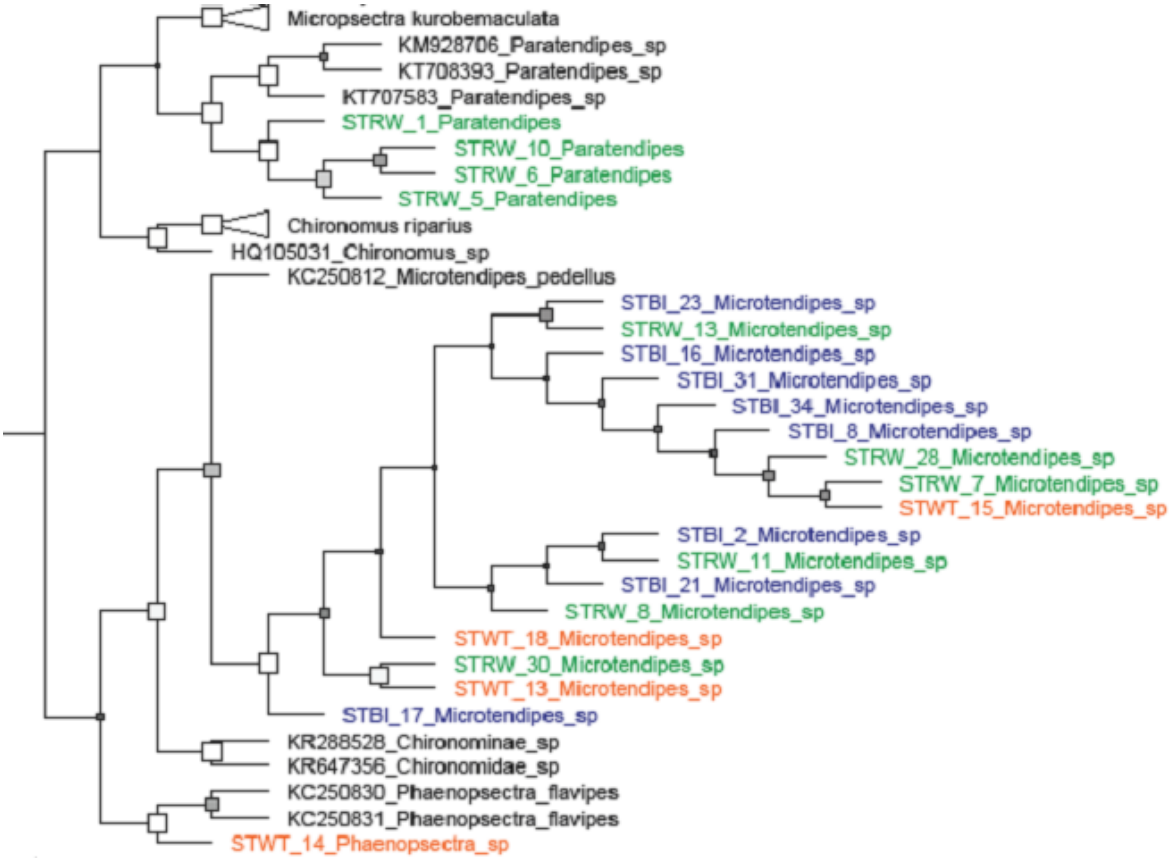


Figure S1 (continued from previous page). Detailed Neighbour-Joining phylogenetic tree of chironomids from the River Stour and GenBank. A) clade containing members of the subfamily Orthocladiinae, and B) clade containing *Paratanytarsus* belonging to the subfamily Chironominae, and C) clade containing other members of the subfamily Chironominae. (See Fig. 3 in main text for full tree.)